

Bayesian Inference with Posterior Regularization and Infinite Latent Support Vector Machines

Jun Zhu

DCSZJ@MAIL.TSINGHUA.EDU.CN

Ning Chen

CHENN07@MAILS.TSINGHUA.EDU.CN

State Key Laboratory of Intelligent Technology and Systems

Tsinghua National Laboratory for Information Science and Technology

Department of Computer Science and Technology

Tsinghua University

Eric P. Xing

EPXING@CS.CMU.EDU

School of Computer Science

Carnegie Mellon University

Editor: ?

Abstract

Existing Bayesian models, especially nonparametric Bayesian methods, rely heavily on specially conceived priors to incorporate domain knowledge for discovering improved latent representations. While priors can affect posterior distributions through Bayes' theorem, imposing posterior regularization is arguably more direct and in some cases can be more natural and easier. In this paper, we present *regularized Bayesian inference* (RegBayes), a computational framework to perform posterior inference with a convex regularization on the desired post-data posterior distributions. RegBayes covers both directed Bayesian networks and undirected Markov networks whose Bayesian formulation results in hybrid chain graph models. When the convex regularization is induced from a linear operator on the posterior distributions, RegBayes can be solved with convex analysis theory. Furthermore, we present two concrete examples of RegBayes, *infinite latent support vector machines* (iLSVM) and *multi-task infinite latent support vector machines* (MT-iLSVM), which explore the large-margin idea in combination with a nonparametric Bayesian model for discovering predictive latent features for classification and multi-task learning, respectively. We present efficient inference methods and report empirical studies on several benchmark datasets, which appear to demonstrate the merits inherited from both large-margin learning and Bayesian nonparametrics. Such results were not available until now, and contribute to push forward the interface between these two important subfields, which have been largely treated as isolated in the community.

Keywords: Bayesian inference, regularization, Bayesian nonparametrics, large-margin learning, classification, multi-task learning

1. Introduction

Bayesian inference, one of the elegant statistical estimation frameworks, is becoming increasingly popular not only in building artificial systems (Pearl, 1988; Bishop, 2006) that handle uncertainties but also in the efforts to develop a theory of how the brain works (Ernst and Banks, 2002; Knill and Pouget, 2004; Tenenbaum et al., 2011). At the core of the Bayesian way

of thinking is the Bayes’ theorem (aka Bayes’ rule), which offers a mathematically rigorous computational mechanism to “reverse engineer” a physical generative process to find the distribution of the hidden structures that likely generated the observed data. Recently, nonparametric Bayesian models have gained remarkable popularity, partly owing to their desirable “nonparametric” nature which allows practitioners to sidestep the difficult model selection problem, e.g., figuring out the unknown number of components (or classes) in a mixture model (Antoniak, 1974) or determining the unknown dimensionality of latent features (Griffiths and Ghahramani, 2005), by using an appropriate prior distribution with a large support. Furthermore, nonparametric Bayesian models allow the model complexity to grow as more data are observed, which is also the key factor that makes nonparametric Bayesian models different from other standard Bayesian models. Among the most commonly used priors are Gaussian process (GP) (Rasmussen and Ghahramani, 2002), Dirichlet process (DP) (Ferguson, 1973; Antoniak, 1974) and Indian buffet process (IBP) (Griffiths and Ghahramani, 2005).

However, standard nonparametric Bayesian models usually make strict and unrealistic assumptions on data, such as that observations being homogeneous or exchangeable. A number of recent developments in Bayesian nonparametrics have attempted to relax such assumptions. For example, to handle heterogeneous observations, predictor-dependent processes (MacEachern, 1999; Williamson et al., 2010) have been proposed; and to relax the exchangeability assumption, various correlation structures, such as hierarchical structures (Teh et al., 2006), temporal or spatial dependencies (Beal et al., 2002; Blei and Frazier, 2010), and stochastic ordering dependencies (Hoff, 2003; Dunson and Peddada, 2007), have been introduced. Although this progress has been substantial, developing sufficiently flexible nonparametric priors still has a long way to go to meet the needs of modeling complex data. Furthermore, almost all the existing nonparametric Bayesian methods rely solely on crafting or learning (Welling et al., 2012) a nonparametric Bayesian prior encoding some special structures¹, which *indirectly* influence the posterior distribution of interest via trading-off with likelihood models through the Bayes’ rule. Since it is the post-data posterior distributions, which capture the latent structures to be learned, that are of our ultimate interest, an arguably more *direct* way to learn a desirable latent-variable model is to impose posterior regularization (i.e., regularization on posterior distributions), as we will explore in this paper. Another reason for using posterior regularization is that in some cases it is more natural and easier to incorporate side domain knowledge or structures, such as the large-margin constraints (Jaakkola et al., 1999; Zhu et al., 2009), constraints defined on a manifold structure (Huh and Fienberg, 2010) or the general expectation constraints defined with various forms of side information (Mann and McCallum, 2010), directly on posterior distributions rather than through priors.

Posterior regularization, usually through imposing constraints on the posterior distributions of latent variables or via some information projection, has been widely studied in learning a finite log-linear model from partially observed data (e.g., semi-supervised learning and learning with side information, such as labeled features), including generalized

1. Although likelihood function is another dimension that can be changed to incorporate domain knowledge, existing work on Bayesian nonparametric methods has been mainly focusing on the prior distributions. Following this convention, this paper assumes that a common likelihood model (e.g., Gaussian likelihood for continuous data) is given.

expectation (Mann and McCallum, 2010), posterior regularization (Ganchev et al., 2010), and alternating projection (Bellare et al., 2009), all of which are doing maximum likelihood estimation (MLE) to learn a single set of model parameters by optimizing an objective that is regularized by posterior constraints. Recent attempts toward learning a posterior distribution of model parameters include the “learning from measurements” (Liang et al., 2009), maximum entropy discrimination (Jaakkola et al., 1999) and MedLDA (maximum entropy discrimination latent Dirichlet allocation) (Zhu et al., 2009). But again, all these methods are restricted to finite parametric models. To the best of our knowledge, very few attempts have been made to impose posterior regularization on nonparametric Bayesian latent variable models.

Technically, although it is intuitively natural for MLE-based methods (i.e., maximizing a likelihood-based objective function with hidden variables) to include a regularization term on the posterior distributions of latent variables when performing an EM-like procedure, this is not straightforward for Bayesian inference using the classic Bayes’ rule because we do not have an optimization objective to be regularized. Although Bayesian inference with hard posterior constraints can be heuristically implemented, e.g., using rejection sampling (Bishop, 2006), it could be extremely inefficient when the sample space is high dimensional. Things will get even worse when the posterior constraints are soft, i.e., allowing some violations but the degree of violation is unknown. Soft constraints could lead to an uncountably many feasible subspaces (each with a different complexity or penalty), which make a rejection sampling method generally infeasible (Please see Figure 1(b) for an illustration).

To offer a mathematically rigorous computational framework for dealing with both hard and soft posterior constraints, in this paper we present a general formulation of *regularized Bayesian inference* (RegBayes), which offers an extra dimension of freedom to standard Bayesian inference by imposing appropriate regularization on the post-data posterior distributions. We base our work on the fresh information theoretical interpretation of the Bayes’ theorem by Zellner (Zellner, 1988), namely, the Bayes’ theorem can be reformulated as a KL-divergence minimization problem. Under this optimization framework, we incorporate posterior constraints to do regularized Bayesian inference, with a penalty term that measures the violation of the constraints. RegBayes covers the broad spectrum of graphical models, including both directed Bayesian networks and undirected Markov networks. For undirected models, the resulting model is a hybrid chain graph (Frydenberg, 1990) when performing Bayesian inference (Murray and Ghahramani, 2004; Qi et al., 2005; Welling and Parise, 2006), which is usually much more challenging than the Bayesian inference in directed Bayesian networks. When the convex regularization is induced from a linear operator (e.g., expectation) of the posterior distributions, RegBayes can be solved with convex analysis theory.

By allowing to use constraints directly on post-data posterior distributions, we believe that the extra flexibility of RegBayes can be beneficial and stimulate new developments in Bayesian nonparametrics and Bayesian inference in general. In this paper, we particularly concentrate on illustrating how to use the ideas of RegBayes to push forward the interface between Bayesian nonparametrics and large margin learning, which have complementary advantages but have been largely treated as two isolated subfields in the community. As the core idea of support vector machines (Vapnik, 1995) and maximum entropy

discrimination (Jaakkola et al., 1999) as well as their structured extensions of max-margin Markov networks (Taskar et al., 2003) and maximum entropy discrimination Markov networks (Zhu and Xing, 2009), large margin learning has shown great success in many scenarios. But a large margin model rarely has the flexibility of nonparametric Bayesian models to automatically resolve model complexity from empirical data, especially when latent variables are present (Jebara, 2001; Zhu et al., 2009). Specifically, we develop the *infinite latent support vector machines* (iLSVM) and *multi-task infinite latent support vector machines* (MT-iLSVM), which explore the discriminative large-margin idea to learn infinite latent feature models for classification and multi-task learning (Argyriou et al., 2007; Bakker and Heskes, 2003), respectively. Both iLSVM and MT-iLSVM are special cases of RegBayes that explore the large-margin principle to consider supervised information for learning predictive latent features, which are good for classification or multi-task learning. For iLSVM, we use the IBP prior to allow the model to have an unbounded number of latent features *a priori*. For MT-iLSVM, we use the similar IBP prior to infer a latent projection matrix to capture the correlations among multiple predictive tasks while avoiding pre-specifying the dimensionality of the projection matrix. The regularized inference problems can be efficiently solved with an iterative procedure, which leverages existing high-performance convex optimization techniques. As a by-product, we also show that MedLDA (Zhu et al., 2009) is a RegBayes model, but with a finite number of latent features.

The rest of the paper is structured as follows. Section 2 discusses related work. Section 3 presents the general framework of regularized Bayesian inference (RegBayes), together with the convex duality results that will be needed in latter sections. Section 4 concretizes the ideas of RegBayes and presents two infinite latent feature models with large-margin constraints for both classification and multi-task learning. Section 5 presents some preliminary experimental results. Finally, Section 6 concludes and discusses future research directions.

2. Related Work

Bayesian inference is one of the most successful paradigms to model uncertainty of empirical data arising in scientific and engineering domains. Bishop (Bishop, 2006) discusses many popular examples in his seminal book, but the book mainly focuses on finite parametric models. Recently, nonparametric Bayesian inference has attracted much attention in statistics and machine learning, and many proposals have been made towards developing a full Bayesian treatment of much richer forms of objects, such as sequential data, grouped data, data with a tree structure and relational data. Gershman and Blei (Gershman and Blei, 2011) presents a nice tutorial on this subject.

Expectation regularization or expectation constraints have also been considered to regularize model parameter estimation in the context of semi-supervised learning or learning with weakly labeled data. Mann and McCallum (Mann and McCallum, 2010) summarizes the recent developments of the generalized expectation (GE) criteria for training a discriminative probabilistic model (e.g., maximum entropy models or conditional random fields (Lafferty et al., 2001)) with unlabeled data. By providing appropriate side information, such as labeled features or estimates of label distributions, a GE-based penalty function is defined to regularize the model distribution, e.g., the distribution of class labels.

One commonly used GE function is the KL-divergence between empirical expectation and model expectation of some feature functions. Although the GE criteria can be used alone as a scoring function to estimate the unknown parameters of a discriminative model, it is more usually used as a regularization term to an estimation method, such as maximum (conditional) likelihood estimation. Bellare et al. (Bellare et al., 2009) presented a different formulation of using expectation constraints in semi-supervised learning by introducing an auxiliary distribution to GE, together with an alternating projection algorithm, which can be more efficient. Liang et al. (Liang et al., 2009) proposed to use the general notion of “measurements” to encapsulate the variety of weakly labeled data for learning an exponential family model. The measurements can be labels, partial labels or other constraints on model predictions. Under the EM framework, posterior constraints are used in (Graca et al., 2009) to modify the E-step of an EM algorithm to project the model posterior distributions onto the subspace of distributions that satisfy a set of auxiliary constraints.

Dudik et al. (Dudik et al., 2007) studies the generalized maximum entropy principle with a rich form of expectation constraints using convex duality theory, where the standard moment matching constraints of maximum entropy are relaxed to inequality constraints. But their analysis was restricted to KL-divergence minimization (maximum entropy principle is a special case) and the finite dimensional space of observations. Later on, Altun and Smola (Altun and Smola, 2006) presents a more general duality theory for a family of divergence functions on Banach spaces. We have drawn a lot of inspiration from both papers to develop the regularized Bayesian inference framework using convex duality theory.

Regularized Bayesian inference provides a computational framework for developing non-parametric Bayesian models with appropriate posterior constraints. The present paper provides a full extension of our preliminary work (Zhu et al., 2011b,a). For example, the infinite SVM (iSVM) (Zhu et al., 2011b) is a latent class model, where each data example is assigned to a single mixture component (i.e., an 1-dimensional space), and both iLSVM and MT-iLSVM extend the ideas to infinite latent feature models. For multi-task learning, non-parametric Bayesian models have been developed in (Xue et al., 2007; Rai and Daume III, 2010) for learning features shared by multiple tasks. However, these methods are based on standard Bayesian inference, without the ability to consider posterior regularization, such as the large-margin constraints or the manifold constraints (Huh and Fienberg, 2010). Finally, MT-iLSVM is a nonparametric Bayesian formulation of the popular multi-task learning methods (Ando and Zhang, 2005; Jebara, 2011).

3. Regularized Bayesian Inference

In this section, we present the computational framework of regularized Bayesian inference. We begin with a brief review of the basic results due to Zellner (Zellner, 1988).

3.1 Bayesian Inference as a Learning Model

Let \mathbb{M} be a space, containing all the random variables of a physical generative process whose posterior distributions we are trying to infer from empirical data. Let us first consider the case of full Bayesian inference, where \mathbb{M} is also the model space and each element $\mathcal{M} \in \mathbb{M}$ represents a model. We will discuss the setting of empirical Bayesian inference shortly, where the model has some unknown model parameters. At the core of Bayesian inference

is Bayes' theorem, which offers a computational procedure to combine prior knowledge and empirical data. More formally, Bayesian inference starts with a prior distribution $\pi(\mathcal{M})$ and a likelihood function $p(\mathbf{x}|\mathcal{M})$ indexed by the model $\mathcal{M} \in \mathbb{M}$. Then, given a collection of observed data $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, the posterior distribution is

$$p(\mathcal{M}|\mathcal{D}) = \frac{\pi(\mathcal{M})p(\mathcal{D}|\mathcal{M})}{p(\mathcal{D})} = \frac{\pi(\mathcal{M})\prod_{n=1}^N p(\mathbf{x}_n|\mathcal{M})}{p(\mathbf{x}_1, \dots, \mathbf{x}_N)}, \quad (1)$$

where $p(\mathcal{D})$ is the marginal likelihood or evidence on observed data. Under the criteria of optimal information processing rule, Zellner (Zellner, 1988) first showed that Bayes' rule is optimal and 100% efficient; and the posterior distribution due to the Bayes' theorem is the same as the optimum solution of the convex variational problem

$$\begin{aligned} \min_{p(\mathcal{M})} \quad & \text{KL}(p(\mathcal{M})\|\pi(\mathcal{M})) - \int_{\mathcal{M}} \log p(\mathcal{D}|\mathcal{M})p(\mathcal{M})d\mathcal{M} \\ \text{s.t. :} \quad & p(\mathcal{M}) \in \mathcal{P}_{\text{prob}}, \end{aligned} \quad (2)$$

where $\text{KL}(p(\mathcal{M})\|\pi(\mathcal{M}))$ is the Kullback-Leibler (KL) divergence, and $\mathcal{P}_{\text{prob}}$ is the space of valid probability distributions with an appropriate dimension. The constraint is due to the law of conservation of belief. Zellner called $p(\mathcal{M})$ a post-data distribution (a pdf for continuous variables) in order to distinguish it from the posterior distribution by Bayes' theorem. Given the equivalence, we will call $p(\mathcal{M})$ a posterior distribution in the sequel if no confusion arises.

As commented by E.T. Jaynes (Zellner, 1988), "this fresh interpretation of Bayes' theorem could make the use of Bayesian methods more attractive and widespread, and stimulate new developments in the general theory of inference". Below, we study how to extend the basic results to incorporate posterior constraints in Bayesian inference.

3.2 Regularized Bayesian Inference with Expectation Constraints

In standard Bayesian inference, although the constraint due to the law of conservation of belief (i.e., $p(\mathcal{M}) \in \mathcal{P}_{\text{prob}}$) does not consider domain knowledge or structures, the above formulation offers one way to extend the scope of Bayesian inference. Formally, we present *regularized Bayesian inference* (RegBayes) as a novel computational procedure to combine prior knowledge and empirical data by solving the constrained optimization problem

$$\begin{aligned} \min_{p(\mathcal{M}), \boldsymbol{\xi}} \quad & \text{KL}(p(\mathcal{M})\|\pi(\mathcal{M})) - \int_{\mathcal{M}} \log p(\mathcal{D}|\mathcal{M})p(\mathcal{M})d\mathcal{M} + U(\boldsymbol{\xi}) \\ \text{s.t. :} \quad & p(\mathcal{M}) \in \mathcal{P}_{\text{post}}(\boldsymbol{\xi}), \end{aligned} \quad (3)$$

where $\mathcal{P}_{\text{post}}(\boldsymbol{\xi})$ is a subspace of distributions that satisfy a set of constraints besides the standard normalization constraints of a probability. To distinguish, we will call a problem *unconstrained* if it only has the standard normalization constraints or does not have any constraints at all.

Although different types of constraints could arise in practice, this paper focuses on the expectation constraints, of which each one is a function of $p(\mathcal{M})$ through an expectation

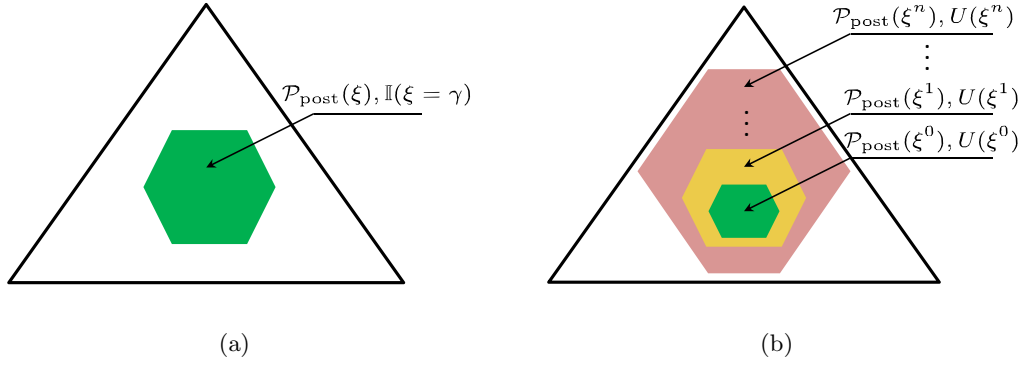


Figure 1: Illustration for the (a) hard and (b) soft constraints in the simple setting which has only three possible models. For hard constraints, we have only one feasible subspace. In contrast, we have many (normally infinite for continuous ξ) feasible subspaces for soft constraints and each of them is associated with a different complexity or penalty, measured by the U function.

operator. For instance, let ψ_t be a feature function defined on \mathcal{M} , a constraint can be of the form

$$h(Ep(\psi_t)) \leq \xi_t, \quad (4)$$

where E is the expectation operator, i.e., $Ep(\psi_t) = \mathbb{E}_{\mathcal{M} \sim p}[\psi_t(\mathcal{M})]$. The auxiliary parameters ξ are usually nonnegative and interpreted as slack variables. The constraints with non-trivial ξ are soft constraints. But we emphasize that by defining U as an indicator function, the formulation (3) covers the case where hard constraints are imposed. For instance, if we define

$$U(\xi) = \sum_t \mathbb{I}(\xi_t = \gamma_t) = \mathbb{I}(\xi = \gamma),$$

where $\mathbb{I}(c)$ is an indicator function that equals to 0 if the condition c is satisfied; otherwise ∞ , then all the expectation constraints (4) are hard constraints. As illustrated in Figure 1(a), hard constraints define one single feasible subspace (assuming to be non-empty). In general, we assume that $U(\xi)$ is a convex function, which measures the complexities of the feasible subspaces, as illustrated in Figure 1(b). A larger subspace typically leads to a higher complexity. In the classification models to be presented, U corresponds to a surrogate loss, e.g., hinge loss of a prediction rule, as we shall see. In fact, the constrained formulation of RegBayes can be equivalently written in an “unconstrained” form

$$\min_{p(\mathcal{M}) \in \mathcal{P}_{\text{prob}}} \text{KL}(p(\mathcal{M}) \| \pi(\mathcal{M})) - \int_{\mathcal{M}} \log p(\mathcal{D} | \mathcal{M}) p(\mathcal{M}) d\mathcal{M} + g(Ep(\mathcal{M})) \quad (5)$$

If we have T features, then the linear operator E (i.e., expectation) maps p to a point in \mathbb{R}^T . We assume that the real-valued function $g : \mathbb{R}^T \rightarrow \mathbb{R}$ is convex and left lower semi-continuous. For each U , we can induce a g function; vice versa. If we use hard constraints,

similar as in regularized maximum entropy density estimation (Altun and Smola, 2006; Dudík et al., 2007), we will have

$$g(Ep) = \sum_t \mathbb{I}(h(Ep(\psi_t)) \leq \gamma_t). \quad (6)$$

For the regularization function g , as well as U , we can have many choices, besides the above mentioned indicator function. For example, if we could obtain empirical expectations of some feature functions $\mathbb{E}_{\tilde{p}}[\psi_t]$ from observed data, one natural regularization function would be the KL-divergence between empirical expectations and the expectations computed from the model distribution, i.e., $g(Ep) = \sum_t \text{KL}(\mathbb{E}_{\tilde{p}}[\psi_t] \| Ep(\psi_t))$ or the generalized Bregman divergence for unnormalized expectations. This regularization function has been used in (Mann and McCallum, 2010) for label regularization, in the context of semi-supervised learning. Other choices include equality constraints, box constraints, and ℓ_2^2 penalty (Please see Table 1 in (Dudík et al., 2007) for a summary). We will also present three new examples shortly for developing latent support vector machines.

3.2.1 GENERALIZATION BEYOND BAYESIAN NETWORKS

Standard Bayesian inference and the proposed RegBayes implicitly make the assumption that the model can be graphically drawn as a Bayesian network as illustrated in Figure 2(a)². Here, we consider a more general formulation which could cover both directed and undirected latent variable models, such as the well-studied Boltzmann machines (Murray and Ghahramani, 2004; Welling et al., 2004), as well as the case where a model could have some unknown parameters (e.g., hyper-parameters) and need an estimation procedure, such as maximum likelihood estimation (MLE), besides posterior inference. The latter is also known as empirical Bayesian methods, which are frequently employed by practitioners.

Extension 1: Empirical Bayesian Inference with Unknown Parameters: As illustrated in Figure 2(b), in some cases we need to perform the empirical Bayesian inference in the presence of unknown parameters. For instance, in a linear-Gaussian Bayesian model, we may choose to estimate its covariance matrix using MLE; and in a latent Dirichlet allocation (LDA) (Blei et al., 2003) model, we may choose to estimate the unknown topical dictionary, although in principle we can treat these parameters as random variables and perform full Bayesian inference. In such cases, we need some mechanisms to estimate the unknown parameters when doing Bayesian inference. Let Θ be model parameters. We can formulate empirical Bayesian inference as solving³

$$\begin{aligned} \min_{\Theta, p(\mathcal{M}|\Theta)} \quad & \text{KL}(p(\mathcal{M}|\Theta) \| \pi(\mathcal{M})) - \int_{\mathcal{M}} \log p(\mathcal{D}|\mathcal{M}, \Theta) p(\mathcal{M}|\Theta) d\mathcal{M} \\ \text{s.t. :} \quad & p(\mathcal{M}|\Theta) \in \mathcal{P}_{\text{prob}}(\Theta). \end{aligned} \quad (7)$$

Although the problem is convex over $p(\mathcal{M}|\Theta)$ for any fixed Θ , it is not jointly convex in general. A natural algorithm to solve this problem is the well-known EM procedure (Dempster et al., 1977), which converges to a local optimum. Specifically, we have the following result.

2. The structure within \mathcal{M} can be arbitrary, either a directed, undirected or hybrid chain graph.
 3. The objective can be derived using variational techniques. It is in fact a variational upper bound of the negative log-likelihood.

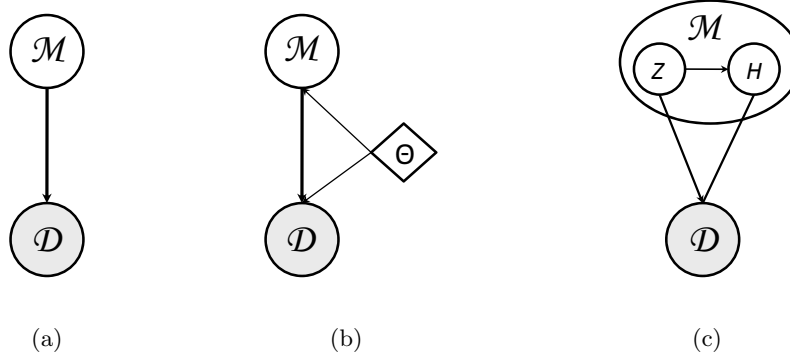


Figure 2: Illustration graphs for three different types of models that involve Bayesian inference: (a) a Bayesian generative model; (b) a Bayesian generative model with unknown parameters Θ ; and (c) a chain graph model.

Lemma 1 *For problem (7), the optimum solution of $p(\mathcal{M}|\Theta)$ is equivalent to the posterior distribution by Bayes' theorem for any Θ ; and the optimum Θ^* is the MLE*

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \log p(\mathcal{D}|\Theta).$$

Proof [Sketch] For any Θ , by the calculus of variation and Lagrangian methods with a Lagrange multiplier ζ , we can get that the optimum solution of $p(\mathcal{M}|\Theta)$ is $p^*(\mathcal{M}|\Theta) = \frac{1}{\exp(1+\zeta)} \pi(\mathcal{M})p(\mathcal{D}|\mathcal{M}, \Theta)$. Due to the normalization constraint, we have $\exp(1 + \zeta) = p(\mathcal{D}|\Theta)$. Thus, $p^*(\mathcal{M}|\Theta)$ is the posterior distribution inferred via the Bayes' theorem. Substituting $p^*(\mathcal{M}|\Theta)$ into the objective of problem (7), we prove the second half. ■

Extension 2: Chain Graph: In the above cases, we have assumed that the observed data are generated by some model in a directed causal sense. This assumption holds in directed latent variable models. However, in many cases, we may choose alternative formulations to define the joint distribution of a model and the observed data. Figure 2(c) illustrates one such scenario, where the model \mathcal{M} consists of two subsets of random variables. One subset H is connected to the observed data via an undirected graph and the other subset Z is connected to the observed data and H using directed edges. This graph is known as a chain graph. Due to the Markov properties of chain graph (Frydenberg, 1990), we know that the joint distribution has the factorization form as

$$p(\mathcal{M}, \mathcal{D}) = p(Z)p(H, \mathcal{D}|Z), \quad (8)$$

where $p(H, \mathcal{D}|Z)$ is a Markov random field (MRF). One concrete example of such a hybrid chain model is the Bayesian Boltzmann machines (Murray and Ghahramani, 2004), which treat the parameters of a Boltzmann machine as random variables and perform Bayesian inference with MCMC sampling methods.

The insights that RegBayes covers undirected or chain graph latent variable models come from the observation that the objective $\mathcal{L}_B(p(\mathcal{M}))$ of problem (2) is in fact an KL-divergence, namely, we can show that

$$\mathcal{L}_B(p(\mathcal{M})) = \text{KL}(p(\mathcal{M}) \| p(\mathcal{M}, \mathcal{D})), \quad (9)$$

where $p(\mathcal{M}, \mathcal{D})$ is the joint distribution. For directed Bayesian networks (Zhu et al., 2011a), we naturally have $p(\mathcal{M}, \mathcal{D}) = \pi(\mathcal{M})p(\mathcal{D}|\mathcal{M})$. For the undirected MRF models, we have $\mathcal{M} = \{Z, H\}$ and again we can define the joint distribution as in Eq. (8).

Putting the above two extensions of Bayesian inference together, the regularized Bayesian inference with estimating unknown model parameters can be generally formulated as

$$\begin{aligned} \min_{\Theta, p(\mathcal{M}|\Theta), \xi} \quad & \mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta)) + U(\xi) \quad \text{or} \quad \min_{\Theta, p(\mathcal{M}|\Theta)} \quad \mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta)) + g(Ep(\mathcal{M})) \quad (10) \\ \text{s.t. :} \quad & p(\mathcal{M}|\Theta) \in \mathcal{P}_{\text{post}}(\Theta, \xi) \quad \quad \quad \text{s.t. :} \quad p(\mathcal{M}|\Theta) \in \mathcal{P}_{\text{prob}}(\Theta), \end{aligned}$$

where $\mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta))$ is the objective function of problem (7). These two formulations are equivalent. We will call the former a *constrained* formulation and call the latter an *unconstrained* formulation by ignoring the standard normalization constraints, which are easy to deal with.

3.2.2 OPTIMIZATION WITH CONVEX DUALITY THEORY

Depending on several factors, including the data likelihood model, the prior and the regularization function, a RegBayes problem in general is highly non-trivial to solve, either in the constrained or unconstrained form. Furthermore, as we have discussed, if Θ is non-empty, the problem is not joint convex, and we need to resort to an iterative procedure. For example, an EM procedure to solve the unconstrained form could be that we iteratively solve for $p(\mathcal{M}|\Theta)$ with Θ fixed; and solve for Θ with $p(\mathcal{M}|\Theta)$ given. The second step can be solved with numerical methods, such as gradient descent. Below, we focus on solving the first step, which is also the whole RegBayes problem for a full Bayesian model (i.e., Θ is null). We know that the first step is convex if U and g are convex. In this section, we present some results of convex analysis theory to deal with the convex RegBayes problem (5) with expectation regularization or the first step of the iterative procedure that solves problem (10).

To make the following statements general, we consider the following problem

$$\min_{x \in \mathcal{X}} f(x) + g(Ax) \quad (11)$$

where $f : \mathcal{X} \rightarrow \mathbb{R}$ is a convex function; $A : \mathcal{X} \rightarrow \mathcal{B}$ is a bounded linear operator; and $g : \mathcal{B} \rightarrow \mathbb{R}$ is also convex. Here, we introduce the convex analysis theory to study this problem by formulating the primal-dual space relations of convex optimization problems in the general settings, where both \mathcal{X} and \mathcal{B} are Banach spaces. One important result is the Fenchel duality theorem.

Definition 2 (Convex Conjugate) *Let \mathcal{X} be a Banach space and \mathcal{X}^* be its dual space. The convex conjugate or the Legendre-Fenchel transformation of a function $f : \mathcal{X} \rightarrow [-\infty, +\infty]$ is $f^* : \mathcal{X}^* \rightarrow [-\infty, +\infty]$, where*

$$f^*(x^*) = \sup_{x \in \mathcal{X}} \{\langle x, x^* \rangle - f(x)\}. \quad (12)$$

Theorem 3 (Fenchel Duality (Borwein and Zhu, 2005)) *Let \mathcal{X} and \mathcal{B} be Banach spaces, $f : \mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ and $g : \mathcal{B} \rightarrow \mathbb{R} \cup \{+\infty\}$ be convex functions and $A : \mathcal{X} \rightarrow \mathcal{B}$ be a bounded linear map. Define the primal and dual values t, d by the Fenchel problems*

$$t = \inf_{x \in \mathcal{X}} \{f(x) + g(Ax)\} \text{ and } d = \sup_{x^* \in \mathcal{B}^*} \{-f^*(A^*x^*) - g^*(-x^*)\}.$$

Then these values satisfy the weak duality inequality $p \geq d$. If f , g and A satisfy either

$$0 \in \text{core}(\text{dom}g - A\text{dom}f) \text{ and both } f \text{ and } g \text{ are left side continuous (lsc),} \quad (13)$$

or

$$A\text{dom}f \cap \text{cont}g \neq \emptyset, \quad (14)$$

then $t = d$ and the supremum to the dual problem is attainable if finite.

The Fenchel duality theorem can be applied to solve divergence minimization problems for density estimation (Altun and Smola, 2006; Dudík et al., 2007). Let $\boldsymbol{\psi} \stackrel{\text{def}}{=} (\psi_1, \dots, \psi_T)$ be a vector of feature functions $\psi_t : \mathcal{X} \rightarrow \mathcal{B}$ and let A be the expectation operator of the feature functions with respect to the distribution p on \mathcal{X} , that is, $Ap \stackrel{\text{def}}{=} \mathbb{E}_{x \sim p}[\boldsymbol{\psi}(x)]$, where $\boldsymbol{\psi}(x) = (\psi_1(x), \dots, \psi_T(x))$. Given a set of observed data $\mathcal{D} = \{x_d\}_{d=1}^D$, we let $\tilde{\boldsymbol{\psi}}$ denote the observed empirical values of the features, namely, $\tilde{\boldsymbol{\psi}} = \frac{1}{D} \sum_{d=1}^D \boldsymbol{\psi}(x_d)$. Then, when the f function is a KL-divergence and the constraints are relaxed moment matching constraints, the following result can be proved.

Lemma 4 (KL-divergence with Constraints (Altun and Smola, 2006))

$$\begin{aligned} & \min_p \left\{ \text{KL}(p||q) \text{ s.t. : } \|\mathbb{E}_p[\boldsymbol{\psi}] - \tilde{\boldsymbol{\psi}}\|_{\mathcal{B}} \leq \epsilon \text{ and } p \in \mathcal{P}_{\text{prob}} \right\} \\ &= \max_{\boldsymbol{\phi}} \left\{ \langle \boldsymbol{\phi}, \tilde{\boldsymbol{\psi}} \rangle - \log \int_{\mathcal{X}} q(x) \exp(\langle \boldsymbol{\phi}, \boldsymbol{\psi}(x) \rangle) dx - \epsilon \|\boldsymbol{\phi}\|_{\mathcal{B}^*} + e^{-1} \right\}, \end{aligned} \quad (15)$$

where the unique solution is given by $\hat{p}_{\hat{\boldsymbol{\phi}}}(x) = q(x) \exp(\langle \hat{\boldsymbol{\phi}}, \boldsymbol{\psi}(x) \rangle - \Lambda_{\hat{\boldsymbol{\phi}}})$ and $\hat{\boldsymbol{\phi}}$ is the solution of the dual problem.

The problem in the above lemma has hard constraints, and the corresponding g is the indicator function $\mathbb{I}(\|\mathbb{E}_p[\boldsymbol{\psi}] - \tilde{\boldsymbol{\psi}}\|_{\mathcal{B}} \leq \epsilon)$ in order to apply the Fenchel duality theorem. Many other examples of the posterior constraints can be found in (Dudík et al., 2007; Mann and McCallum, 2010), as we have discussed in Section 3.2. In this paper, we consider the general soft constraints as in the RegBayes problem (3). Furthermore, we do not assume the existence of a fully observed dataset to compute the empirical expectation $\tilde{\boldsymbol{\phi}}$. Specifically, we have the following result.

Lemma 5 (RegBayes) *Let E be the expectation operator with feature functions $\boldsymbol{\psi}$ defined on \mathcal{M} , and assume g is convex. We have*

$$\begin{aligned} & \min_{p(\mathcal{M})} \left\{ \text{KL}(p(\mathcal{M})||p(\mathcal{M}, \mathcal{D})) + g(Ep) \text{ s.t. : } p(\mathcal{M}) \in \mathcal{P}_{\text{prob}} \right\} \\ &= \max_{\boldsymbol{\phi}} \left\{ -\log \int_{\mathcal{M}} p(\mathcal{M}, \mathcal{D}) \exp(\langle \boldsymbol{\phi}, \boldsymbol{\psi}(\mathcal{M}) \rangle) d\mathcal{M} - g^*(-\boldsymbol{\phi}) \right\}, \end{aligned} \quad (16)$$

where the unique solution is given by $\hat{p}_{\hat{\phi}}(\mathcal{M}) = p(\mathcal{M}, \mathcal{D}) \exp(\langle \hat{\phi}, \psi(\mathcal{M}) \rangle - \Lambda_{\hat{\phi}})$ and $\hat{\phi}$ is the solution of the dual problem.

Now, we derive the conjugate functions of three other important examples, which will be used shortly for developing the infinite latent SVM models. We defer the proof to Appendix. Specifically, the first one is the conjugate of a simple function, which will be used in a binary latent SVM classification model.

Lemma 6 *Let $g_0 : \mathbb{R} \rightarrow \mathbb{R}$ be defined as $g_0(x) = C \max(0, x)$. Then, we have*

$$g_0^*(\mu) = \mathbb{I}(0 \leq \mu \leq C).$$

The second function is slightly more complex, which will be used for defining a multi-way latent SVM classification model. Specifically, let $\mathcal{G} = \{\mathbf{x} \in \mathbb{R}^L : \exists j, x_j \geq 0\}$. We define the function $g_1 : \mathcal{G} \rightarrow \mathbb{R}$ as

$$g_1(\mathbf{x}) = C \max(\mathbf{x}), \quad (17)$$

where $\max(\mathbf{x}) \stackrel{\text{def}}{=} \max(x_1, \dots, x_L)$. Apparently, g_1 is convex because it is a point-wise maximum (Boyd and Vandenberghe, 2004) of the simple linear functions $\phi_i(\mathbf{x}) = x_i$. Then, we have the following results.

Lemma 7 *The convex conjugate of $g_1(\mathbf{x})$ as defined above is*

$$g_1^*(\boldsymbol{\mu}) = \mathbb{I}\left(\forall i, \mu_i \geq 0; \text{ and } \sum_j \mu_j \leq C\right).$$

Let $\mathcal{G}' \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^2 : x_1 + x_2 = 0\}$ and (c_1, c_2) are constants. The last function that we are interested in is $g_2 : \mathcal{G}' \rightarrow \mathbb{R}$, where

$$g_2(\mathbf{x}; c_1, c_2) = C(\max(0, x_1 - c_1) + \max(0, x_2 - c_2)). \quad (18)$$

Then, we have the following lemma, which will be used in developing large-margin regression model.

Lemma 8 *The convex conjugate of $g_2(\mathbf{x})$ as defined above is*

$$g_2^*(\boldsymbol{\mu}; c_1, c_2) = (c_1\mu_1 + c_2\mu_2) + \mathbb{I}\left(\forall i, 0 \leq \mu_i \leq C; \text{ and } \mu_1\mu_2 = 0\right).$$

Note that although g_2 and g_2^* are defined in two dimensional spaces, the feature values of \mathbf{x} or $\boldsymbol{\mu}$ are in fact lying in lower (i.e., 1) dimensional subspaces because of the constraints.

3.3 MedLDA: A RegBayes Model with Finite Latent Features

Before we present the nonparametric regularized Bayesian models, which could have an unbounded number of hidden units, we end this section with a new interpretation of the previously proposed MedLDA (maximum entropy discrimination latent Dirichlet allocation) model (Zhu et al., 2009) under the framework of regularized Bayesian inference. In MedLDA, each data example is projected to a point in a finite dimensional latent space, of

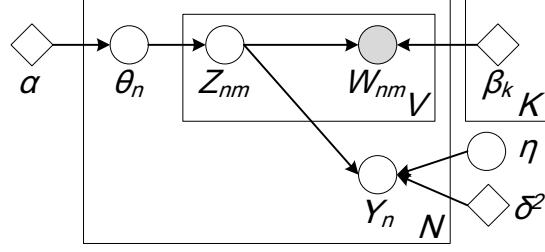


Figure 3: Graphical structure of MedLDA.

which each feature corresponds to a topic, i.e., a unigram distribution over the terms in a vocabulary. MedLDA represents each data as a probability distribution over the features, which results in a conservation constraint (i.e., the more a data expresses on one feature, the less it can express others) (Griffiths and Ghahramani, 2005). The infinite latent feature models discussed later do not have such a constraint.

Without loss of generality, we consider the MedLDA regression model as an example (classification model is similar), whose graphical structure is shown in Figure 3, where we have assumed all data examples have the same length V for notation simplicity. Let K be the number of topics or the dimensionality of the latent topic space. Define $\bar{Z}_n = \frac{1}{V} \sum_{m=1}^V Z_{nm}$ and let $\Theta = \{\alpha, \beta, \delta^2\}$ denote the unknown model parameters and $\mathcal{D} = \{y_n, w_{nm}\}$ be the training set. MedLDA was defined as solving a regularized MLE problem with expectation constraints

$$\begin{aligned} \min_{\xi, \xi^*} \quad & -\log p(\{y_n, w_{nm}\}|\Theta) + C \sum_{n=1}^N (\xi_n + \xi_n^*) \\ \text{s.t. } \forall n : \quad & \begin{cases} y_n - \mathbb{E}_p[\boldsymbol{\eta}^\top \bar{Z}_n] \leq \epsilon + \xi_n \\ -y_n + \mathbb{E}_p[\boldsymbol{\eta}^\top \bar{Z}_n] \leq \epsilon + \xi_n^* \\ \xi_n, \xi_n^* \geq 0 \end{cases} \end{aligned} \quad (19)$$

The posterior constraints are imposed following the large-margin principle and they correspond to a quality measure of the prediction results on training data. In fact, it is easy to show that minimizing $U(\boldsymbol{\xi}, \boldsymbol{\xi}^*) = C \sum_{n=1}^N (\xi_n + \xi_n^*)$ under the above constraints is equivalent to minimizing an ϵ -insensitive loss (Smola and Schölkopf, 2003)

$$\mathcal{R}_\epsilon(p(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\mathcal{D}, \Theta)) = C \sum_{n=1}^N \max(0, |y_n - \mathbb{E}_p[\boldsymbol{\eta}^\top \bar{Z}_n]| - \epsilon). \quad (20)$$

of the expected linear prediction rule $\hat{y}_n = \mathbb{E}_p[\boldsymbol{\eta}^\top \bar{Z}_n]$. Therefore, MedLDA can be seen as belonging to the framework of the GE criteria (Mann and McCallum, 2010), but in the context of large-margin learning.

To practically learn an MedLDA model, since the above problem is intractable, variational methods were used by introducing an auxiliary distribution $q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta)$ ⁴ to approximate the true posterior $p(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\mathcal{D}, \Theta)$, replacing the negative data likelihood

4. We have explicitly written the condition on model parameters.

with its upper bound $\mathcal{L}(q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta))$, and replacing p by q in the constraints. The variational MedLDA regression model is

$$\begin{aligned} \min_{q, \Theta, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \quad & \mathcal{L}(q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta)) + C \sum_{n=1}^N (\xi_n + \xi_n^*) \\ \text{s.t. } \forall n : \quad & \begin{cases} y_n - \mathbb{E}_q[\boldsymbol{\eta}^\top \bar{Z}_n] \leq \epsilon + \xi_n \\ -y_n + \mathbb{E}_q[\boldsymbol{\eta}^\top \bar{Z}_n] \leq \epsilon + \xi_n^* \\ \xi_n, \xi_n^* \geq 0 \end{cases} \end{aligned} \quad (21)$$

where $\mathcal{L}(q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta)) = -\mathbb{E}_q[\log p(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}, \mathcal{D}|\Theta)] - \mathcal{H}(q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta))$ is a variational upper-bound of the negative data log-likelihood. Note that the upper bound is tight if no restricting constraints are made on the variational distribution q . In practice, additional assumptions (e.g., mean-field) can be made on q to derive a practical approximate algorithm.

Based on the previous discussions on the extensions of RegBayes and the duality in Lemma 1, we can reformulate the MedLDA regression model as an example of RegBayes. Specifically, for the MedLDA regression model, we have $\mathcal{M} = \{\theta_n, z_{nm}, \boldsymbol{\eta}\}$. According to Eq. (9), we can easily show that

$$\begin{aligned} \mathcal{L}(q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta)) &= \text{KL}(q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta) \| p(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}, \{w_{nm}, y_n\}|\Theta)) \\ &= \mathcal{L}_B(\Theta, q(\mathcal{M}|\Theta)). \end{aligned}$$

Then, the MedLDA problem is a RegBayes model in Eq. (10) with

$$\mathcal{P}_{\text{post}}^{\text{MedLDA}}(\Theta, \boldsymbol{\xi}, \boldsymbol{\xi}^*) \stackrel{\text{def}}{=} \left\{ q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta) \left| \begin{array}{l} \forall n : \quad y_n - \mathbb{E}_q[\boldsymbol{\eta}^\top \bar{Z}_n] \leq \epsilon + \xi_n \\ \quad \quad -y_n + \mathbb{E}_q[\boldsymbol{\eta}^\top \bar{Z}_n] \leq \epsilon + \xi_n^* \\ \quad \quad \xi_n, \xi_n^* \geq 0 \end{array} \right. \right\}. \quad (22)$$

For the MedLDA problem, we can use Lagrangian methods to solve the constrained formulation. Alternatively, we can also use the convex duality theorem to solve the equivalent unconstrained form. For the variational MedLDA, the ϵ -insensitive loss is $\mathcal{R}_\epsilon(q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta))$. Its conjugate can be derived using the results of Lemma 8. Specifically, we have the following result, whose proof is deferred to Appendix A.4.

Lemma 9 (Conjugate of MedLDA) *For the variational MedLDA problem, we have*

$$\begin{aligned} & \min_{\Theta, q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta) \in \mathcal{P}_{\text{prob}}} \mathcal{L}(q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta), \Theta) + \mathcal{R}_\epsilon(q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta)) \\ &= \max_{\boldsymbol{\omega}} -\log Z'(\boldsymbol{\omega}, \Theta^*) - \sum_n g_2^*(\boldsymbol{\omega}_n; -y_n + \epsilon, y_n + \epsilon), \end{aligned} \quad (23)$$

where $\boldsymbol{\omega}_n = (\omega_n, \omega'_n)$. Moreover, The optimum distribution is the posterior distribution

$$\hat{q}(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta^*) = \frac{1}{Z'(\hat{\boldsymbol{\omega}}, \Theta^*)} p(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}, \mathcal{D}|\Theta^*) \exp \left\{ \sum_n (\omega_n - \omega'_n) \boldsymbol{\eta}^\top \bar{z}_n \right\}, \quad (24)$$

where $Z'(\hat{\boldsymbol{\omega}}, \Theta)$ is the normalization factor and the optimum parameters are

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \log p(\mathcal{D}|\Theta). \quad (25)$$

Note that although in general, either the primal or the dual problem is hard to solve exactly, the above conjugate results are still useful when developing approximate inference algorithms. For instance, we can impose additional mean-field assumptions on q in the primal formulation and iteratively solve for each factor; and in this process convex conjugates are useful to deal with the large-margin constraints (Zhu et al., 2009). Alternatively, we can apply approximate methods (e.g., MCMC sampling) to infer the q based on its solution in Eq. (24), and iteratively solves for the dual parameters ω using approximate statistics (Schofield, 2006). We will discuss more on this when presenting the inference algorithms for iLSVM and MT-iLSVM.

In the above discussions, we have treated the topics β as fixed unknown parameters. A fully Bayesian formulation would treat β as random variables, e.g., with a Dirichlet distribution prior as in (Blei et al., 2003; Griffiths and Steyvers, 2004). Under the RegBayes interpretation, we can easily do such an extension of MedLDA, simply by moving β from Θ to \mathcal{M} .

4. Infinite Latent Support Vector Machines

As we have stated above, MedLDA is a RegBayes model which has a finite number of latent features (i.e., topics) and the dimensionality is pre-specified. In this section, we present two nonparametric RegBayes models to illustrate how to develop latent large-margin classifiers and automatically resolve the unknown dimensionality of latent features from data. We consider two settings. For single-task classification, we consider learning latent features that can be used as a representation of examples to make prediction, and for multi-task learning, we consider learning a common latent projection matrix that captures relationships among the multiple tasks.

We first present the single-task classification model. The basic setup is that we project each data example $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^D$ to a latent feature vector \mathbf{z} . Here, we consider binary features⁵. Given a set of N data examples, let \mathbf{Z} be the matrix, of which each row is a binary vector \mathbf{z}_n associated with data sample n . Instead of pre-specifying a fixed dimension of \mathbf{z} , we resort to the nonparametric Bayesian methods and let \mathbf{z} have an infinite number of dimensions. To make the expected number of active latent features finite, we put the well-studied IBP prior on the binary feature matrix \mathbf{Z} .

4.1 Indian Buffet Process

Indian buffet process (IBP) was proposed in (Griffiths and Ghahramani, 2005) and has been successfully applied in various fields, such as link prediction (Miller et al., 2009) and multi-task learning (Rai and Daume III, 2010). We focus on its stick-breaking construction (Teh et al., 2007), which is good for developing efficient inference methods. Let $\pi_k \in (0, 1)$ be a parameter associated with each column of the binary matrix \mathbf{Z} . Given π_k , each z_{nk} in column k is sampled independently from $\text{Bernoulli}(\pi_k)$. The parameter π are

5. Real-valued features can be easily considered as in (Griffiths and Ghahramani, 2005).

generated by a stick-breaking process

$$\pi_1 = \nu_1, \text{ and } \pi_k = \nu_k \pi_{k-1} = \prod_{i=1}^k \nu_i, \quad (26)$$

where $\nu_i \sim \text{Beta}(\alpha, 1)$. This process results in a decreasing sequence of π_k . Specifically, given a finite dataset, the probability of seeing feature k decreases exponentially with k .

4.2 Infinite Latent Support Vector Machines

We consider the multi-way classification, where each training data is provided with a categorical label y , where $y \in \mathcal{Y} \stackrel{\text{def}}{=} \{1, \dots, L\}$. For binary classification and regression, similar procedure can be applied to impose large-margin constraints on posterior distributions. Suppose that the latent features \mathbf{z} are given, then we can define the *latent discriminant function* as linear

$$f(y, \mathbf{x}, \mathbf{z}; \boldsymbol{\eta}) \stackrel{\text{def}}{=} \boldsymbol{\eta}^\top \mathbf{g}(y, \mathbf{x}, \mathbf{z}), \quad (27)$$

where $\mathbf{g}(y, \mathbf{x}, \mathbf{z})$ is a vector stacking L subvectors⁶ of which the y th is \mathbf{z}^\top and all the others are zero. Since we are doing Bayesian inference, we need to maintain the entire distribution profile of the latent features \mathbf{Z} . However, in order to make a prediction on the observed data \mathbf{x} , we need to remove the uncertainty of \mathbf{Z} . Here, we define the *effective discriminant function* as an expectation⁷ (i.e., a weighted average considering all possible values of \mathbf{Z}) of the latent discriminant function. To make the model fully Bayesian, we also treat $\boldsymbol{\eta}$ as random and aim to infer its posterior distribution from given data. More formally, the effective discriminant function $f : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ is

$$f(y, \mathbf{x}; p(\mathbf{Z}, \boldsymbol{\eta})) \stackrel{\text{def}}{=} \mathbb{E}_{p(\mathbf{Z}, \boldsymbol{\eta})} [f(y, \mathbf{x}, \mathbf{z}; \boldsymbol{\eta})] = \mathbb{E}_{p(\mathbf{Z}, \boldsymbol{\eta})} [\boldsymbol{\eta}^\top \mathbf{g}(y, \mathbf{x}, \mathbf{z})], \quad (28)$$

where $p(\mathbf{Z}, \boldsymbol{\eta})$ is the posterior distribution we want to infer.

With the above definitions, we define the $\mathcal{P}_{\text{post}}(\boldsymbol{\xi})$ in problem (3) using soft⁸ large-margin constraints as

$$\mathcal{P}_{\text{post}}^c(\boldsymbol{\xi}) \stackrel{\text{def}}{=} \left\{ p(\mathbf{Z}, \boldsymbol{\eta}) \left| \begin{array}{l} \forall n \in \mathcal{I}_{\text{tr}} : f(y_n, \mathbf{x}_n; p(\mathbf{Z}, \boldsymbol{\eta})) - f(y, \mathbf{x}_n; p(\mathbf{Z}, \boldsymbol{\eta})) \geq \ell_n^\Delta(y) - \xi_n, \forall y \\ \xi_n \geq 0 \end{array} \right. \right\}$$

and define the penalty function as

$$U^c(\boldsymbol{\xi}) \stackrel{\text{def}}{=} C \sum_{n \in \mathcal{I}_{\text{tr}}} \xi_n^p,$$

6. We can consider the input features \mathbf{x} or its certain statistics in combination with the latent features \mathbf{z} to define a classifier boundary, by simply concatenating them in the subvectors.

7. Although other choices such as taking the mode are possible, our choice could lead to a computationally easy problem because expectation is a linear functional of the distribution under which the expectation is taken. Moreover, expectation can be more robust than taking the mode (Khan et al., 2010), and it has been widely used in (Zhu et al., 2009, 2011b).

8. Hard constraints for the separable cases are covered by simply setting $\boldsymbol{\xi} = 0$.

where $p \geq 1$. If p is 1, minimizing $U^c(\boldsymbol{\xi})$ is equivalent to minimizing the hinge-loss (or ℓ_1 -loss) \mathcal{R}_h^c of the prediction rule (35), where

$$\mathcal{R}_h^c = C \sum_{n \in \mathcal{I}_{\text{tr}}} \max_y (f(y, \mathbf{x}_n; p(\mathbf{Z}, \boldsymbol{\eta})) + \ell_n^\Delta(y) - f(y_n, \mathbf{x}_n; p(\mathbf{Z}, \boldsymbol{\eta})));$$

if p is 2, the surrogate loss is the ℓ_2 -loss. For clarity, we consider the hinge loss. The non-negative cost function $\ell_n^\Delta(y)$ (e.g., 0/1-cost) measures the cost of predicting \mathbf{x}_n to be y when its true label is y_n . \mathcal{I}_{tr} is the index set of training data and I is an identity matrix with appropriate dimensions.

In order to robustly estimate the latent matrix \mathbf{Z} , we need a reasonable amount of data. Therefore, we also relate \mathbf{Z} to the observed data \mathbf{x} by defining a likelihood model to provide as much data as possible. Here, we define the most common linear-Gaussian likelihood model for real-valued data

$$p(\mathbf{x}_n | \mathbf{z}_n, \mathbf{W}, \sigma_{n0}^2) = \mathcal{N}(\mathbf{x}_n | \mathbf{W} \mathbf{z}_n^\top, \sigma_{n0}^2 I), \quad (29)$$

where \mathbf{W} is a random loading matrix. We assume \mathbf{W} follows an independent Gaussian prior, i.e., $\pi(\mathbf{W}) = \prod_d \mathcal{N}(\mathbf{w}_d | 0, \sigma_0^2 I)$. Figure 4 (a) shows the graphical structure of iLSVM. The hyperparameters σ_0^2 and σ_{n0}^2 can be set a priori or estimated from observed data (See Appendix A.7 for details).

Training: Putting the above definitions together, we get the RegBayes problem for iLSVM in the following two equivalent forms

$$\begin{aligned} \min_{p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}), \boldsymbol{\xi}} \quad & \text{KL}(p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) \| p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}, \mathcal{D})) + U^c(\boldsymbol{\xi}) \\ \text{s.t. :} \quad & p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) \in \mathcal{P}_{\text{post}}^c \end{aligned} \quad (30)$$

$$\iff \min_{p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) \in \mathcal{P}_{\text{prob}}} \text{KL}(p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) \| p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}, \mathcal{D})) + \mathcal{R}_h^c(p(\mathbf{Z}, \boldsymbol{\eta})) \quad (31)$$

Here, $p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) \in \mathcal{P}_{\text{post}}^c$ means that the marginal distribution $p(\mathbf{Z}, \boldsymbol{\eta})$ belongs to $\mathcal{P}_{\text{post}}^c$.

Note that in order to be a valid RegBayes model, we need to ensure that the objective function and the posterior constraints have finite values. This can be verified as follows. Although the number of latent features is allowed to be infinite, with probability one, the number of non-zero features is finite when only a finite number of data are observed, under the IBP prior. Moreover, because of the facts that the KL-term in Eq. (3) has the “zero forcing” property (Bishop, 2006, Chap. 10) and the prior distribution of feature z_{nk} decreases exponentially as k increases, we can expect that the posterior distribution of feature z_{nk} also decreases exponentially, when a finite set of data is observed. Thus, both the objective function and the large-margin constraints are well-defined. Finally, to make the problem computationally feasible, we usually set a finite upper bound K to the number of possible features, where K is sufficiently large and known as the truncation level (See Section 4.4 and Appendix A.7 for details). As shown in (Doshi-Velez et al., 2009), the ℓ_1 -distance truncation error of marginal distributions decreases exponentially as K increases.

Directly solving the iLSVM problems is not easy because either the posterior constraints or the non-smooth regularization function \mathcal{R}^c is hard to deal with. Thus, we resort to convex

duality theory, which will be useful for developing approximate inference algorithms, as we have discussed in Section 3.3. We can either solve the constrained form (30) using Lagrangian duality theory (Ito and Kunisch, 2008) or solve the unconstrained form (31) using Fenchel duality theory. Here, we take the second approach. In this case, the linear operator is the expectation operator, denoted by $E : \mathcal{P}_{\text{prob}} \rightarrow \mathbb{R}^{D \times L}$ and the element of Ep evaluated at y for the d th example is

$$Ep(n, y) \stackrel{\text{def}}{=} f(y_n, \mathbf{x}_n; p(\mathbf{Z}, \boldsymbol{\eta})) - f(y, \mathbf{x}_n; p(\mathbf{Z}, \boldsymbol{\eta})) = \mathbb{E}_{p(\mathbf{Z}, \boldsymbol{\eta})} [\boldsymbol{\eta}^\top \Delta \mathbf{g}_n(y, \mathbf{Z})], \quad (32)$$

where $\Delta \mathbf{g}_n(y, \mathbf{Z}) \stackrel{\text{def}}{=} \mathbf{g}(y_n, \mathbf{x}_n, \mathbf{z}) - \mathbf{g}(y, \mathbf{x}_n, \mathbf{z})$. We can easily prove that $\forall n, \max_y (\ell_n^\Delta(y) - Ep(n, y)) \geq 0$. Then, let $g_1 : \mathbb{R}^L \rightarrow \mathbb{R}$ be a function defined in the same form as in Eq. (17). We have

$$\mathcal{R}_h^c(p(\mathbf{Z}, \boldsymbol{\eta})) = \sum_n g_1(\ell_n^\Delta - Ep(n)),$$

where $Ep(n) \stackrel{\text{def}}{=} (Ep(n, 1), \dots, Ep(n, L))$ and $\ell_n^\Delta \stackrel{\text{def}}{=} (\ell_n^\Delta(1), \dots, \ell_n^\Delta(L))$ are the vectors of elements evaluated for n th data. By the Fenchel's duality theorem and the results in Lemma 7, we can derive the conjugate of the problem (31). The proof is deferred to Appendix A.5.

Lemma 10 (Conjugate of iLSVM) *For the iLSVM problem, we have that*

$$\begin{aligned} & \min_{p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) \in \mathcal{P}_{\text{prob}}} \text{KL}(p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) \| p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}, \mathcal{D})) + \mathcal{R}_h^c(q(\mathbf{Z}, \boldsymbol{\eta})) \\ &= \max_{\boldsymbol{\omega}} -\log Z(\boldsymbol{\omega}) + \sum_n \sum_y \zeta_n^y \ell_n^\Delta(y) - \sum_n g_1^*(\boldsymbol{\omega}_n), \end{aligned} \quad (33)$$

where $\boldsymbol{\omega}_n = (\omega_n^1, \dots, \omega_n^L)$ is the subvector associated with data n . Moreover, The optimum distribution is the posterior distribution

$$\hat{p}(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) = \frac{1}{Z(\hat{\boldsymbol{\omega}})} p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}, \mathcal{D}) \exp \left\{ \sum_n \sum_y \hat{\omega}_n^y \boldsymbol{\eta}^\top \Delta \mathbf{g}_n(y, \mathbf{Z}) \right\}, \quad (34)$$

where $Z(\hat{\boldsymbol{\omega}})$ is the normalization factor and $\hat{\boldsymbol{\omega}}$ is the solution of the dual problem.

Testing: to make prediction on test examples, we put both training and test data together to do the regularized Bayesian inference. For training data, we impose the above large-margin constraints because of the awareness of their true labels, while for test data, we do the inference without the large-margin constraints since we do not know their true labels. After inference, we make the prediction via the rule

$$y^* \stackrel{\text{def}}{=} \underset{y}{\operatorname{argmax}} f(y, \mathbf{x}; p(\mathbf{Z}, \boldsymbol{\eta})). \quad (35)$$

The ability to generalize to test data relies on the fact that all the data examples share $\boldsymbol{\eta}$ and the IBP prior. We can also cast the problem as a transductive inference problem by imposing additional constraints on test data (Joachims, 1999). However, the resulting problem will be generally harder to solve.

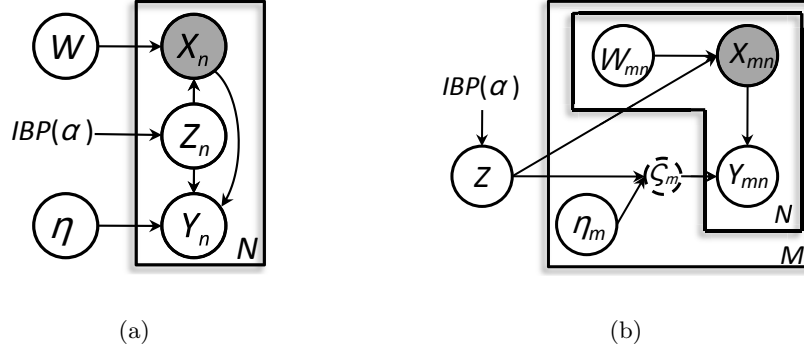


Figure 4: Graphical structures of (a) infinite latent SVM (iLSVM); and (b) multi-task infinite latent SVM (MT-iLSVM). For MT-iLSVM, the dashed nodes (i.e., ς_m) illustrate the task relatedness but do not exist.

4.3 Multi-Task Infinite Latent Support Vector Machines

Different from classification, which is typically formulated as a single learning task, multi-task learning aims to improve a set of related tasks through sharing statistical strength between these tasks, which are performed jointly. Many different approaches have been developed for multi-task learning (See (Jebara, 2011) for a review). In particular, learning a common latent representation shared by all the related tasks has proven to be an effective way to capture task relationships (Ando and Zhang, 2005; Argyriou et al., 2007; Rai and Daume III, 2010). Below, we present the multi-task infinite latent SVM (MT-iLSVM) for learning a common binary projection matrix \mathbf{Z} to capture the relationships among multiple tasks. Similar as in iLSVM, we also put the IBP prior on \mathbf{Z} to allow it to have an unbounded number of columns.

Suppose we have M related tasks. Let $\mathcal{D}_m = \{(\mathbf{x}_{mn}, y_{mn})\}_{n \in \mathcal{I}_{tr}^m}$ be the training data for task m . We consider binary classification tasks, where $\mathcal{Y}_m = \{+1, -1\}$. Extension to multi-way classification or regression can be easily done. Figure 5(a) shows the naïve way that performs multiple tasks independently. In order to make the multiple tasks coupled and share statistical strength, MT-iLSVM introduces a latent projection matrix \mathbf{Z} . If the latent matrix \mathbf{Z} is given, we define the latent discriminant function for task m as

$$f_m(\mathbf{x}_{mn}, \mathbf{Z}; \boldsymbol{\eta}_m) \stackrel{\text{def}}{=} (\mathbf{Z}\boldsymbol{\eta}_m)^\top \mathbf{x}_{mn} = \boldsymbol{\eta}_m^\top (\mathbf{Z}^\top \mathbf{x}_{mn}), \quad (36)$$

where \mathbf{x}_{mn} is one data example in \mathcal{D}_m . This definition provides two views of how the M tasks get related.

- (1) If we let $\varsigma_m = \mathbf{Z}\boldsymbol{\eta}_m$, then ς_m is the actual parameter of task m and all ς_m in different tasks are coupled by sharing the same latent matrix \mathbf{Z} , as illustrated in Figure 5(b);
- (2) Another view is that each task m has its own parameters $\boldsymbol{\eta}_m$, but all the tasks share the same latent projection matrix \mathbf{Z} to extract latent features $\mathbf{Z}^\top \mathbf{x}_{mn}$, which is a projection of the input features \mathbf{x}_{mn} , as illustrated in Figure 5(c).

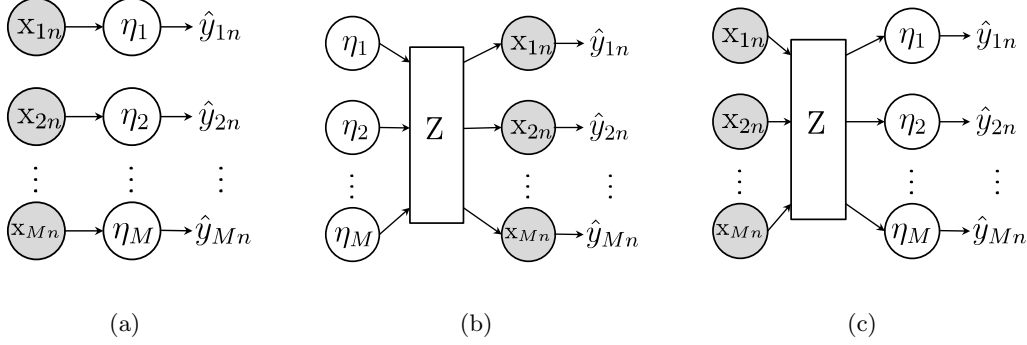


Figure 5: Illustration of (a) multiple single task learning, where each task m (represented by the model η_m) is performed independently; (b) related multiple tasks in MT-iLSVM with the first type of representation, where all the M models need to pass a common transformation (denoted by the matrix \mathbf{Z}) in order to act on input data; and (c) related multiple tasks in MT-iLSVM with the second type of representation, where input data are projected into latent representations using the same projection matrix \mathbf{Z} .

As such, our method can be viewed as a nonparametric Bayesian treatment of alternating structure optimization (ASO) (Ando and Zhang, 2005), which learns a single projection matrix with a pre-specified latent dimension. Moreover, different from (Jebara, 2011), which learns a binary vector with known dimensionality to select features or kernels on \mathbf{x} , we learn an unbounded projection matrix \mathbf{Z} using nonparametric Bayesian techniques.

As in iLSVM, we take the fully Bayesian point of view and treat η_m as random and define the effective discriminant function for task m as the expectation

$$f_m(\mathbf{x}; p(\mathbf{Z}, \eta)) \stackrel{\text{def}}{=} \mathbb{E}_{p(\mathbf{Z}, \eta)} [f_m(\mathbf{x}, \mathbf{Z}; \eta_m)] = \mathbb{E}_{p(\mathbf{Z}, \eta)} [\mathbf{Z} \eta_m]^\top \mathbf{x}. \quad (37)$$

Then, the prediction rule for task m is naturally $y_m^* \stackrel{\text{def}}{=} \text{sign} f_m(\mathbf{x})$. Similarly, we do regularized Bayesian inference by imposing the following constraints and defining

$$U^{MT}(\xi) \stackrel{\text{def}}{=} C \sum_{m, n \in \mathcal{I}_{\text{tr}}^m} \xi_{mn}$$

and

$$\mathcal{P}_{\text{post}}^{MT}(\xi) \stackrel{\text{def}}{=} \left\{ p(\mathbf{Z}, \eta) \left| \begin{array}{l} \forall m, \forall n \in \mathcal{I}_{\text{tr}}^m : y_{mn} \mathbb{E}_{p(\mathbf{Z}, \eta)} [\mathbf{Z} \eta_m]^\top \mathbf{x}_{mn} \geq 1 - \xi_{mn} \\ \xi_{mn} \geq 0 \end{array} \right. \right\}. \quad (38)$$

Finally, to obtain more data to estimate the latent \mathbf{Z} , we also relate it to observed data by defining the likelihood model

$$p(\mathbf{x}_{mn} | \mathbf{w}_{mn}, \mathbf{Z}, \lambda_{mn}^2) = \mathcal{N}(\mathbf{x}_{mn} | \mathbf{Z} \mathbf{w}_{mn}, \lambda_{mn}^2 I), \quad (39)$$

where \mathbf{w}_{mn} is a vector. We assume \mathbf{W} has an independent prior $\pi(\mathbf{W}) = \prod_{mn} \mathcal{N}(\mathbf{w}_{mn}|0, \sigma_{m0}^2 I)$. Fig. 4 (b) illustrates the graphical structure of MT-iLSVM.

For training, we can derive the similar convex conjugate as in the case of iLSVM. Similar as in iLSVM, minimizing $U^{MT}(\boldsymbol{\xi})$ is equivalent to minimizing the hinge-loss \mathcal{R}_h^{MT} of the multiple binary prediction rules, where

$$\mathcal{R}_h^{MT}(q(\mathbf{Z}, \boldsymbol{\eta})) = C \sum_{m,n \in \mathcal{I}_{\text{tr}}^m} \max(0, 1 - y_{mn} \mathbb{E}_{p(\mathbf{Z}, \boldsymbol{\eta})}[\mathbf{Z}\boldsymbol{\eta}_m]^\top \mathbf{x}_{mn}). \quad (40)$$

Thus, the RegBayes problem of MT-iLSVM can be equivalently written as

$$\min_{p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W})} \text{KL}(p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) \| p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}, \mathcal{D})) + \mathcal{R}_h^{MT}(q(\mathbf{Z}, \boldsymbol{\eta})). \quad (41)$$

Then, by the Fenchel's duality theorem and Lemma 6, we can derive the conjugate of MT-iLSVM. The proof is deferred to Appendix A.6.

Lemma 11 (Conjugate of MT-iLSVM) *For the MT-iLSVM problem, we have that*

$$\begin{aligned} & \min_{p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) \in \mathcal{P}_{\text{prob}}} \text{KL}(p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) \| p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}, \mathcal{D})) + \mathcal{R}_h^{MT}(q(\mathbf{Z}, \boldsymbol{\eta})) \\ &= \max_{\boldsymbol{\omega}} -\log Z'(\boldsymbol{\omega}) + \sum_{m,n} \omega_{mn} - \sum_{m,n} g_0^*(\omega_{mn}). \end{aligned} \quad (42)$$

Moreover, The optimum distribution is the posterior distribution

$$\hat{p}(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) = \frac{1}{Z'(\hat{\boldsymbol{\omega}})} p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}, \mathcal{D}) \exp \left\{ \sum_{m,n} y_{mn} \hat{\omega}_{mn} (\mathbf{Z}\boldsymbol{\eta}_m)^\top \mathbf{x}_{mn} \right\}, \quad (43)$$

where $Z'(\hat{\boldsymbol{\omega}})$ is the normalization factor and $\hat{\boldsymbol{\omega}}$ is the solution of the dual problem.

For testing, we use the same strategy as in iLSVM to do Bayesian inference on both training and test data. The difference is that training data are subject to large-margin constraints, while test data are not. Similarly, the hyper-parameters σ_{m0}^2 and λ_{mn}^2 can be set a priori or estimated from data (See Appendix A.7 for details).

4.4 Inference with Truncated Mean-Field Constraints

We discuss how to do regularized Bayesian inference (3) with the large-margin constraints for both iLSVM and MT-iLSVM. From the primal-dual formulations, it is obvious that there are basically two methods to perform the regularized Bayesian inference. One is to directly solve the posterior distribution $p(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W})$, and the other is to first solve the dual problem for the optimum $\hat{\boldsymbol{\omega}}$ and then infer the posterior distribution. However, both the primal and dual problems are intractable to solve for iLSVM and MT-iLSVM. The intrinsic hardness is due to the mutual dependency among the latent variables in the desired posterior distribution. Therefore, a natural approximation method is the mean field (Jordan et al., 1999), which breaks the mutual dependency by assuming p is of some factorization form. This method approximates the original problems by imposing additional constraints. An alternative method is to apply approximate methods (e.g., MCMC sampling) to infer the

Algorithm 1 Inference Algorithm for Infinite Latent SVMs

- 1: **Input:** corpus \mathcal{D} and constants (α, C) .
 - 2: **Output:** posterior distribution $p(\boldsymbol{\nu}, \mathbf{Z}, \boldsymbol{\eta}, \mathbf{W})$.
 - 3: **repeat**
 - 4: infer $p(\boldsymbol{\nu}), p(\mathbf{W})$ and $p(\mathbf{Z})$ with $p(\boldsymbol{\eta})$ and $\boldsymbol{\omega}$ given;
 - 5: infer $p(\boldsymbol{\eta})$ and solve for $\boldsymbol{\omega}$ with $p(\mathbf{Z})$ given.
 - 6: **until** convergence
-

true posterior distributions derived via convex conjugates as above, and iteratively estimate the dual parameters using approximate statistics (e.g., feature expectations estimated using samples) (Schofield, 2006). Below, we use MT-iLSVM as an example to illustrate the idea of the first strategy. A full discussion on the second strategy is beyond the scope of this paper. For iLSVM, similar procedure applies and we defer its details to Appendix A.8.

To make the problem easier to solve, we use the stick-breaking representation of IBP, which includes the auxiliary variable $\boldsymbol{\nu}$, and infer the expanded posterior $p(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta})$. The joint model distribution is now $p(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta}, \mathcal{D})$. Furthermore, we impose the truncated mean-field constraint that

$$p(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta}) = p(\boldsymbol{\eta}) \prod_{k=1}^K \left(p(\nu_k | \gamma_k) \prod_{d=1}^D p(z_{dk} | \psi_{dk}) \right) \prod_{mn} p(\mathbf{w}_{mn} | \Phi_{mn}, \sigma_{mn}^2 I), \quad (44)$$

where K is the truncation level, and we assume that

$$\begin{aligned} p(\nu_k | \gamma_k) &= \text{Beta}(\gamma_{k1}, \gamma_{k2}), \\ p(z_{dk} | \psi_{dk}) &= \text{Bernoulli}(\psi_{dk}), \\ p(\mathbf{w}_{mn} | \Phi_{mn}, \sigma_{mn}^2 I) &= \mathcal{N}(\mathbf{w}_{mn} | \Phi_{mn}, \sigma_{mn}^2 I). \end{aligned}$$

Then, we can use the duality theory⁹ to solve the RegBayes problem by alternating between two substeps, as outlined in Algorithm 1 and detailed below.

Infer $p(\boldsymbol{\nu})$, $p(\mathbf{W})$ and $p(\mathbf{Z})$: Since $p(\boldsymbol{\nu})$ and $p(\mathbf{W})$ are not directly involved in the posterior constraints, we can solve for them by using standard Bayesian inference, i.e., minimizing a KL-divergence. Specifically, for $p(\mathbf{W})$, since the prior is also normal, we can easily derive the update rules for Φ_{mn} and σ_{mn}^2 . For $p(\boldsymbol{\nu})$, we have the same update rules as in (Doshi-Velez et al., 2009). We defer the details to Appendix A.7.

For $p(\mathbf{Z})$, it is directly involved in the posterior constraints. So, we need to solve it together with $p(\boldsymbol{\eta})$ using conjugate theory. However, this is intractable. Here, we adopt an alternating strategy that first infers $p(\mathbf{Z})$ with $p(\boldsymbol{\eta})$ and dual parameters $\boldsymbol{\omega}$ fixed, and then infers $p(\boldsymbol{\eta})$ and solves for $\boldsymbol{\omega}$. Specifically, since the large-margin constraints are linear of $p(\mathbf{Z})$, we can get the mean-field update equation as

$$\psi_{dk} = \frac{1}{1 + e^{-\vartheta_{dk}}},$$

9. Lagrangian duality (Ito and Kunisch, 2008) was used in (Zhu et al., 2011a) to solve the constrained variational formulations, which is closely related to Fenchel duality (Magnanti, 1974) and leads to the same solutions for iLSVM and MT-iLSVM.

where

$$\begin{aligned} \vartheta_{dk} = & \sum_{j=1}^k \mathbb{E}_p[\log v_j] - \mathcal{L}_k^\nu - \sum_{mn} \frac{1}{2\lambda_{mn}^2} \left((K\sigma_{mn}^2 + (\phi_{mn}^k)^2) \right. \\ & \left. - 2x_{mn}^d \phi_{mn}^k + 2 \sum_{j \neq k} \phi_{mn}^j \phi_{mn}^k \psi_{dj} \right) + \sum_{m,n \in \mathcal{I}_{tr}^m} y_{mn} \mathbb{E}_p[\eta_{mk}] x_{mn}^d, \end{aligned} \quad (45)$$

and \mathcal{L}_k^ν is a lower bound of $\mathbb{E}_p[\log(1 - \prod_{j=1}^k v_j)]$ (See Appendix A.7 for details). The last term of ϑ_{dk} is due to the large-margin posterior constraints as defined in Eq. (38). We can how the large-margin constraints regularize the procedure of inferring the latent matrix \mathbf{Z} .

Infer $p(\boldsymbol{\eta})$ and solve for $\boldsymbol{\omega}$: Now, we can apply the convex conjugate theory and show that the optimum posterior distribution of $\boldsymbol{\eta}$ is

$$p(\boldsymbol{\eta}) = \prod_m p(\boldsymbol{\eta}_m), \text{ where } p(\boldsymbol{\eta}_m) \propto \pi(\boldsymbol{\eta}_m) \exp\{\boldsymbol{\eta}_m^\top \boldsymbol{\mu}_m\},$$

and $\boldsymbol{\mu}_m = \sum_{n \in \mathcal{I}_{tr}^m} y_{mn} \omega_{mn} (\boldsymbol{\psi}^\top \mathbf{x}_{mn})$. Here, we assume $\pi(\boldsymbol{\eta}_m)$ is standard normal. Then, we have $p(\boldsymbol{\eta}_m) = \mathcal{N}(\boldsymbol{\eta}_m | \boldsymbol{\mu}_m, I)$ and the optimum dual parameters can be obtained by solving the following M independent dual problems

$$\begin{aligned} \max_{\boldsymbol{\omega}_m} \quad & -\frac{1}{2} \boldsymbol{\mu}_m^\top \boldsymbol{\mu}_m + \sum_{n \in \mathcal{I}_{tr}^m} \omega_{mn} \\ \forall n \in \mathcal{I}_{tr}^m, \text{ s.t. : } \quad & 0 \leq \omega_{mn} \leq C, \end{aligned} \quad (46)$$

where the constraints are from the conjugate function g_0^* in Lemma 11. These dual problems (or their primal forms) can be efficiently solved with a binary SVM solver, such as SVM-light or LibSVM.

5. Experiments

We present empirical results for both classification and multi-task learning. Our results appear to demonstrate the merits inherited from both Bayesian nonparametrics and large-margin learning.

5.1 Multi-way Classification

We evaluate the infinite latent SVM (iLSVM) for classification on the real TRECVID2003 and Flickr image datasets, which have been extensively evaluated in the context of learning finite latent feature models (Chen et al., 2010). TRECVID2003 consists of 1078 video key-frames, and each example has two types of features – 1894-dimension binary vector of text features and 165-dimension HSV color histogram. The Flickr image dataset consists of 3411 natural scene images about 13 types of animals, including *squirrel*, *cow*, *cat*, *zebra*, *tiger*, *lion*, *elephant*, *whales*, *rabbit*, *snake*, *antlers*, *hawk* and *wolf*, downloaded from the Flickr website¹⁰. Also, each example has two types of features, including 500-dimension SIFT bag-of-words and 634-dimension real-valued features (e.g., color histogram, edge direction histogram, and block-wise color moments). Here, we consider the real-valued features only by using Gaussian likelihood distributions for \mathbf{x} .

10. <http://www.flickr.com/>

Table 1: Classification accuracy and F1 scores on the TRECVID2003 and Flickr image datasets.

Model	TRECVID2003		Flickr	
	Accuracy	F1 score	Accuracy	F1 score
EFH+SVM	0.565 ± 0.0	0.427 ± 0.0	0.476 ± 0.0	0.461 ± 0.0
MMH	0.566 ± 0.0	0.430 ± 0.0	0.538 ± 0.0	0.512 ± 0.0
IBP+SVM	0.553 ± 0.013	0.397 ± 0.030	0.500 ± 0.004	0.477 ± 0.009
iLSVM	0.563 ± 0.010	0.448 ± 0.011	0.533 ± 0.005	0.510 ± 0.010

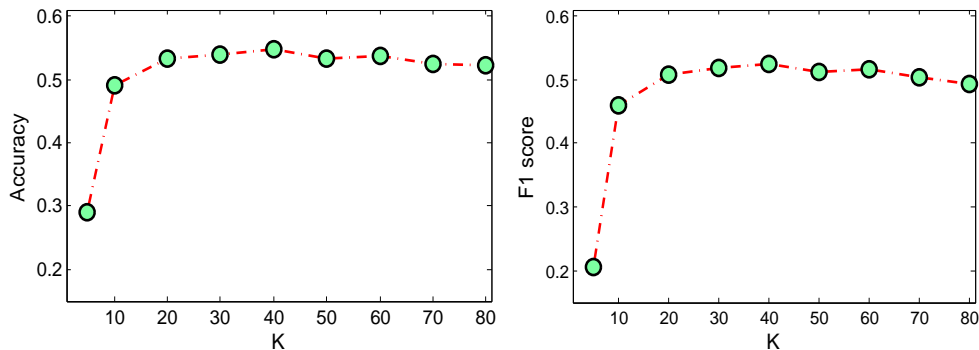


Figure 6: Accuracy and F1 score of MMH on the Flickr dataset with different numbers of latent features.

We compare iLSVM with the large-margin Harmonium (MMH) (Chen et al., 2010), which was shown to outperform many other latent feature models, and two decoupled approaches – *EFH+SVM* and *IBP+SVM*. *EFH+SVM* uses the exponential family Harmonium (EFH) (Welling et al., 2004) to discover latent features and then learns a multi-way SVM classifier. *IBP+SVM* is similar, but uses an IBP factor analysis model (Griffiths and Ghahramani, 2005) to discover latent features. Both MMH and *EFH+SVM* are finite models and they need to pre-specify the dimensionality of latent features. We report their results on classification accuracy and F1 score (i.e., the average F1 score over all possible classes) (Zhu et al., 2011b) achieved with the best dimensionality in Table 1. Figure 6 illustrates the performance change of MMH when using different number of latent features, from which we can see that $K = 40$ produces the best performance and either increasing or decreasing K could make the performance worse. For iLSVM and *IBP+SVM*, we use the mean-field inference method and present the average performance with 5 randomly initialized runs (Please see Appendix A.8 for the algorithm and initialization details). We perform 5-fold cross-validation on training data to select hyperparameters, e.g., α and C (we use the same procedure for MT-iLSVM). We can see that iLSVM can achieve comparable performance with

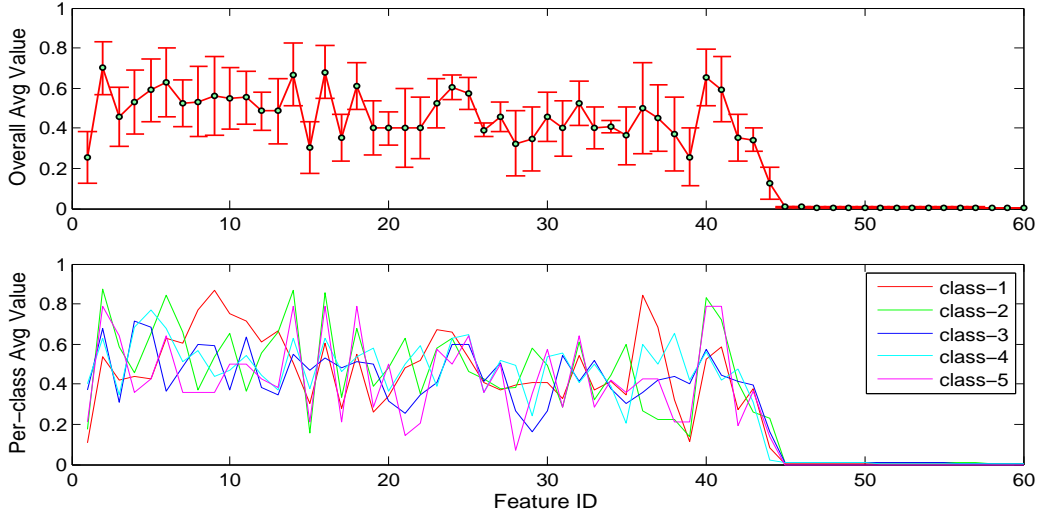


Figure 7: (Up) the overall average values of the latent features with standard deviation over different classes; and (Bottom) the per-class average values of latent features learned by iLSVM on the TRECVID dataset.

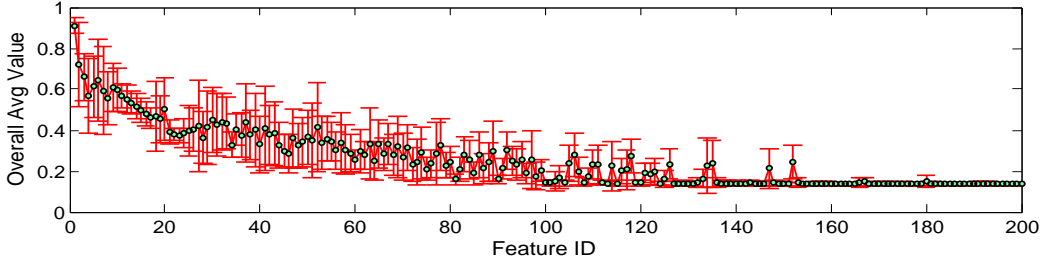


Figure 8: The overall average values of the latent features with standard deviation over different classes on the Flickr dataset.

the nearly optimal MMH, without needing to pre-specify the latent feature dimension¹¹, and is much better than the decoupled approaches (i.e., IBP+SVM and EFH+SVM).

It is also interesting to examine the discovered latent features. Figure 7 shows the overall average values of latent features and the per-class average feature values of iLSVM in one run on the TRECVID dataset. We can see that on average only about 45 features are active for the TRECVID dataset. For the overall average, we also present the standard deviation over the 5 categories. A larger deviation means that the corresponding feature is more discriminative when predicting different categories. For example, feature 26 and feature 34 are generally less discriminative than many other features, such as feature 1 and feature 30. Figure 8 shows the overall average feature values together with standard deviation on the Flickr dataset. We omitted the per-class average because that figure is too

¹¹. We set the truncation level to 300, which is large enough.

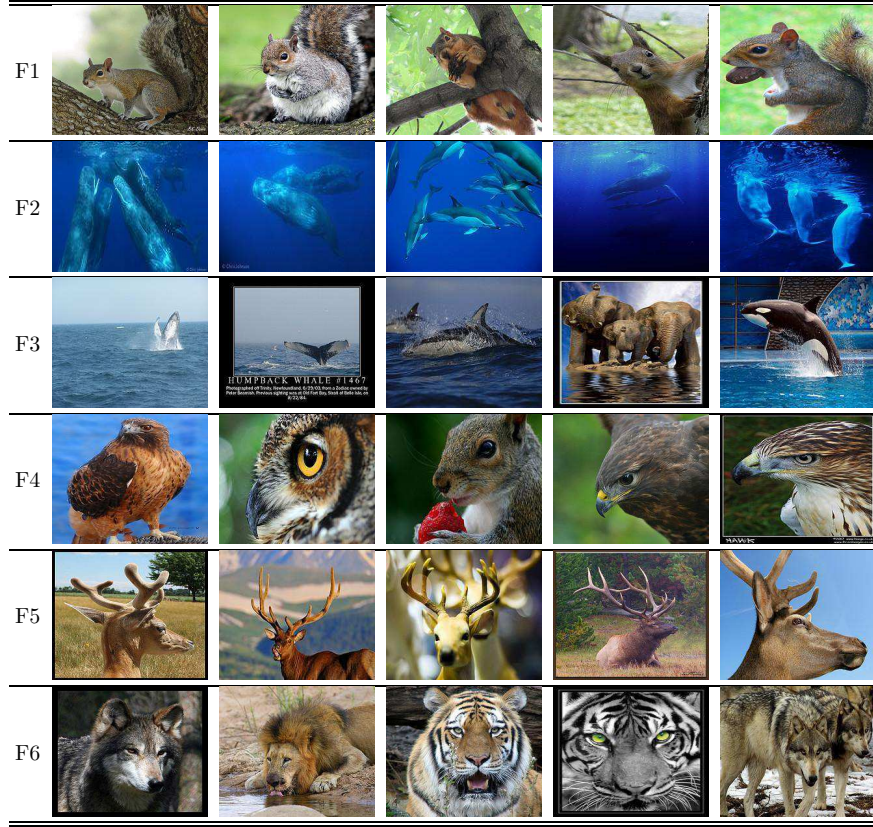


Figure 9: Six example features discovered iLSVM on the Flickr animal dataset. For each feature, we show 5 top-ranked images.

crowded with 13 categories. We can see that as k increases, the probability that feature k is active decreases. The reason for the features with stable values (i.e., standard deviations are extremely small) is due to our initialization strategy (each feature has 0.5 probability to be active). Initializing ψ_{dk} as being exponentially decreasing (e.g., like the constructing process of π) leads to a faster decay and many features will be inactive. To examine the semantics¹² of each feature, Figure 9 presents some example features discovered on the Flickr animal dataset. For each feature, we present 5 top-ranked images which have large values on this particular feature. We can see that most of the features are semantically interpretable. For instance, feature F1 is about squirrel; feature F2 is about ocean animal, which is whales in the Flickr dataset; and feature F4 is about hawk. We can also see that some features are about different aspects of the same category. For example, feature F2 and feature F3 are both about whales, but with different background.

5.2 Multi-task Learning

Now, we evaluate the multi-task infinite latent SVM (MT-iLSVM) on several well-studied real datasets.

12. The interpretation of latent features depends heavily on the input data.

Table 2: Multi-label classification performance on Scene and Yeast datasets.

Dataset	Model	Acc	F1-Micro	F1-Macro
Yeast	YaXue	0.5106	0.3897	0.4022
	Piyushrai-1	0.5212	0.3631	0.3901
	Piyushrai-2	0.5424	0.3946	0.4112
	MT-IBP+SVM	0.5475 ± 0.005	0.3910 ± 0.006	0.4345 ± 0.007
	MT-iLSVM	0.5792 ± 0.003	0.4258 ± 0.005	0.4742 ± 0.008
Scene	YaXue	0.7765	0.2669	0.2816
	Piyushrai-1	0.7756	0.3153	0.3242
	Piyushrai-2	0.7911	0.3214	0.3226
	MT-IBP+SVM	0.8590 ± 0.002	0.4880 ± 0.012	0.5147 ± 0.018
	MT-iLSVM	0.8752 ± 0.004	0.5834 ± 0.026	0.6148 ± 0.020

5.2.1 DESCRIPTION OF THE DATA

Scene and Yeast Data: These datasets are from the UCI repository, and each data example has multiple labels. As in (Rai and Daume III, 2010), we treat the multi-label classification as a multi-task learning problem, where each label assignment is treated as a binary classification task. The Yeast dataset consists of 1500 training and 917 test examples, each having 103 features, and the number of labels (or tasks) per example is 14. The Scene dataset consists 1211 training and 1196 test examples, each having 294 features, and the number of labels (or tasks) per example for this dataset is 6.

School Data: This dataset comes from the Inner London Education Authority and has been used to study the effectiveness of schools. It consists of examination records from 139 secondary schools in years 1985, 1986 and 1987. It is a random 50% sample with 15362 students. The dataset is publicly available and has been extensively evaluated in various multi-task learning methods (Bakker and Heskes, 2003; Bonilla et al., 2008; Zhang and Yeung, 2010), where each task is defined as predicting the exam scores of students belonging to a specific school based on four student-dependent features (year of the exam, gender, VR band and ethnic group) and four school-dependent features (percentage of students eligible for free school meals, percentage of students in VR band 1, school gender and school denomination). In order to compare with the above methods, we follow the same setup described in (Argyriou et al., 2007; Bakker and Heskes, 2003) and similarly we create dummy variables for those features that are categorical forming a total of 19 student-dependent features and 8 school-dependent features. We use the same 10 random splits¹³ of the data, so that 75% of the examples from each school (task) belong to the training set and 25% to the test set. On average, the training set includes about 80 students per school and the test set about 30 students per school.

13. Available at: <http://ttic.uchicago.edu/~argyriou/code/index.html>

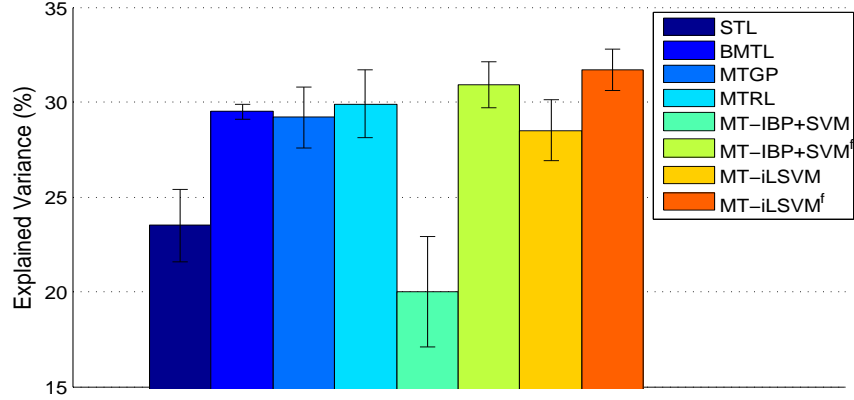


Figure 10: Percentage of explained variance by various models on the School dataset.

5.2.2 RESULTS

Scene and Yeast Data: We compare with the closely related nonparametric Bayesian methods, including kernel stick-breaking (YaXue) (Xue et al., 2007) and the basic and augmented infinite predictor subspace models (i.e., Piyushrai-1 and Piyushrai-2) (Rai and Daume III, 2010). These nonparametric Bayesian models were shown to outperform the independent Bayesian logistic regression and a single-task pooling approach (Rai and Daume III, 2010). We also compare with a decoupled method $MT-IBP+SVM$ ¹⁴ that uses an IBP factor analysis model to find shared latent features among multiple tasks and then builds separate SVM classifiers for different tasks. For MT-iLSVM and MT-IBP+SVM, we use the mean-field inference method in Sec 4.4 and report the average performance with 5 randomly initialized runs (See Appendix A.7 for initialization details). For comparison with (Rai and Daume III, 2010; Xue et al., 2007), we use the overall classification accuracy, F1-Macro and F1-Micro as performance measures. Table 2 shows the results. On both datasets, MT-iLSVM needs less than 50 latent features on average. We can see that the large-margin MT-iLSVM performs much better than other nonparametric Bayesian methods and MT-IBP+SVM, which separates the inference of latent features from learning the classifiers.

School Data: We use the percentage of explained variance (Bakker and Heskes, 2003) as the measure of the regression performance, which is defined as the total variance of the data minus the sum-squared error on the test set as a percentage of the total variance. Since we use the same settings, we can compare with the state-of-the-art results of

- (1) Bayesian multi-task learning (BMTL) (Bakker and Heskes, 2003);
- (2) Multi-task Gaussian processes (MTGP) (Bonilla et al., 2008);
- (3) Convex multi-task relationship learning (MTRL) (Zhang and Yeung, 2010);

and single-task learning (STL) as reported in (Bonilla et al., 2008; Zhang and Yeung, 2010). For MT-iLSVM and MT-IBP+SVM, we also report the results achieved by using both the

14. This decoupled approach is in fact an one-iteration MT-iLSVM, where we first infer the shared latent matrix \mathbf{Z} and then learn an SVM classifier for each task.

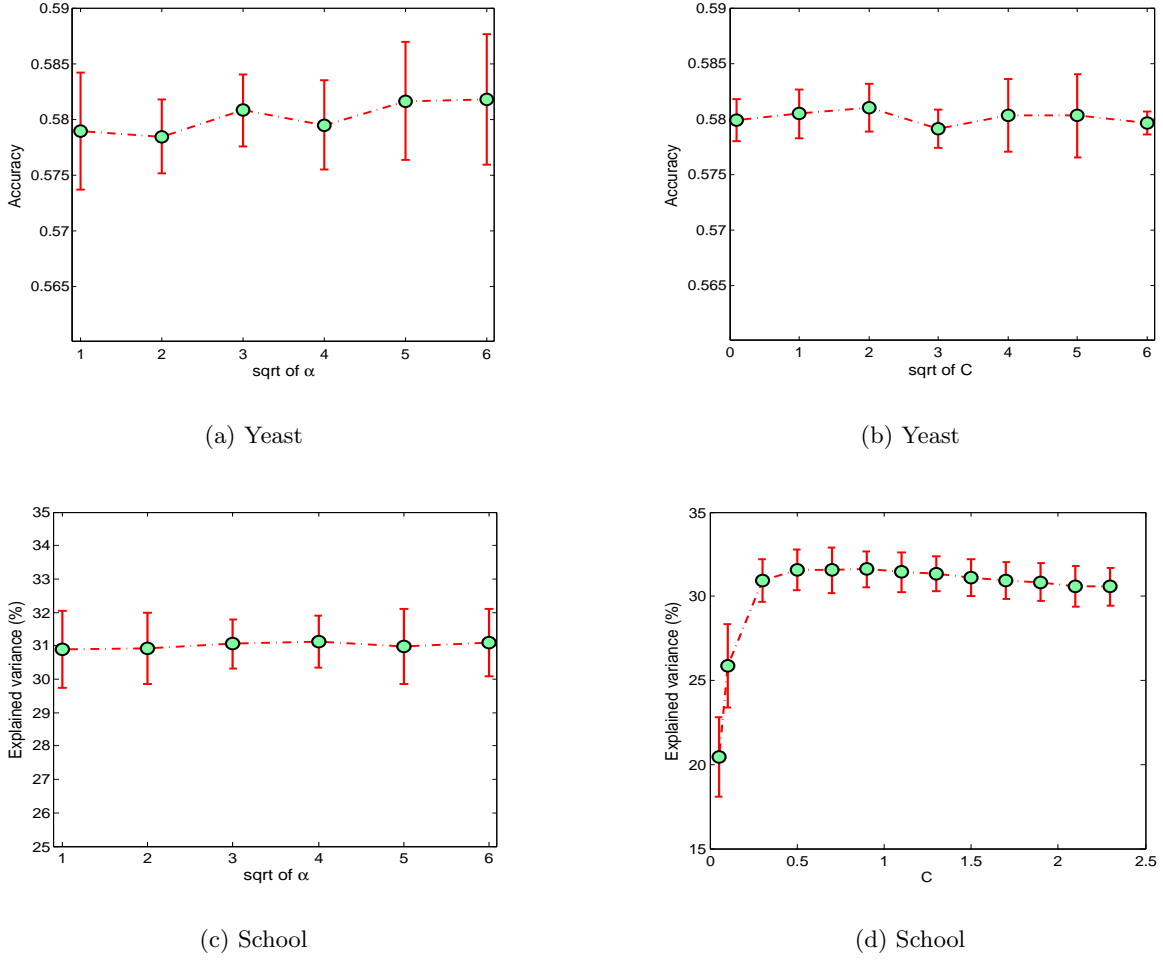


Figure 11: Sensitivity study of MT-iLSVM: (a) classification accuracy with different α on Yeast data; (b) classification accuracy with different C on Yeast data; (c) percentage of explained variance with different α on School data; and (d) percentage of explained variance with different C on School data.

latent features (i.e., $\mathbf{Z}^\top \mathbf{x}$) and the original input features \mathbf{x} through vector concatenation, and we denote the corresponding methods by $MT-iLSVM^f$ and $MT-IBP+SVM^f$, respectively. On average the multi-task latent SVM (i.e., MT-iLSVM) needs about 50 latent features to get sufficiently good and robust performance. From the results in Figure 10, we can see that the MT-iLSVM achieves better results than the existing methods that have been tested in previous studies. Again, the joint MT-iLSVM performs much better than the decoupled method MT-IBP+SVM, which separates the latent feature inference from the training of large-margin classifiers. Finally, using both latent features and the original input features can boost the performance slightly for MT-iLSVM, while much more significantly for the decoupled MT-IBP+SVM.

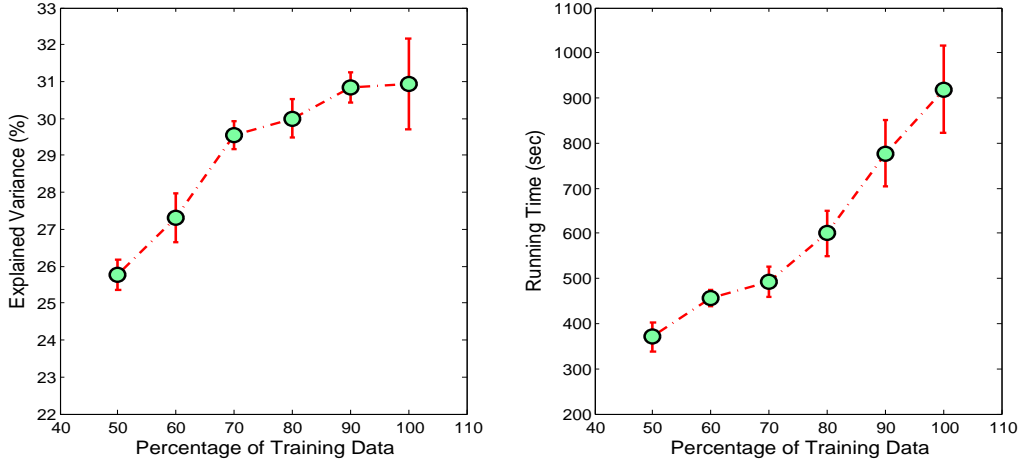


Figure 12: Percentage of explained variance and running time by MT-iLSVM with various training sizes.

5.3 Sensitivity Analysis

Figure 11 shows how the performance of MT-iLSVM changes against the hyper-parameter α and regularization constant C on the Yeast and School datasets. We can see that on the Yeast dataset, MT-iLSVM is insensitive to both α and C . For the School dataset, MT-iLSVM is very insensitive to the α , and it is stable when C is set between 0.3 and 1.

Figure 12 shows how the training size affects the performance and running time of MT-iLSVM on the School dataset. We use the first $b\%$ ($b = 50, 60, 70, 80, 90, 100$) of the training data in each of the 10 random splits as training set and use the corresponding test data as test set. We can see that as training size increases, the performance and running time generally increase; and MT-iLSVM achieves the state-of-art performance when using about 70% training data. From the running time, we can also see that MT-iLSVM is generally quite efficient by using mean-field inference.

Finally, we investigate how the performance of MT-iLSVM changes against the hyper-parameters σ_{m0}^2 and λ_{mn}^2 . We initially set $\sigma_{m0}^2 = 1$ and compute λ_{mn}^2 from observed data. If we further estimate them by maximizing the objective function, the performance does not change much ($\pm 0.3\%$ for average explained variance on the School dataset). We have similar observations for iLSVM.

6. Conclusions and Discussions

We present regularized Bayesian inference (RegBayes), a computational framework to perform post-data posterior inference with a convex regularization on the desired posterior distributions. RegBayes is applicable to both directed and undirected graphical models. General conjugate results are derived when the posterior regularization is induced from a linear operator (e.g., expectation). Furthermore, we particularly concentrate on developing two large-margin nonparametric Bayesian models under the RegBayes framework to

learn predictive latent features for classification and multi-task learning, by exploring the large-margin principle to define posterior constraints. Both models allow the latent dimension to be automatically resolved from the data. The empirical results on several real datasets appear to demonstrate that our methods inherit the merits from both Bayesian nonparametrics and large-margin learning.

Regularized Bayesian inference offers a computational framework for considering posterior regularization in performing nonparametric Bayesian inference. For future work, we plan to study other posterior regularization beyond the large-margin constraints, such as posterior constraints defined on manifold structures (Huh and Fienberg, 2010), and investigate how posterior regularization can be used in other interesting nonparametric Bayesian models (Beal et al., 2002; Teh et al., 2006; Blei and Frazier, 2010) in different contexts, such as link prediction (Miller et al., 2009) for social network analysis. As we have stated, Reg-Bayes can be developed for undirected MRFs. But the inference would be even harder. We plan to do a systematic investigation along this direction. We have some preliminary results presented in (Chen et al., 2011).

Acknowledgements

Jun is supported by National Key Project for Basic Research of China (No. 2012CB316300) and the 221 Basic Research Plan for Young Faculties at Tsinghua University. Eric Xing is supported by AFOSR FA95501010247, ONR N000140910758, NSF Career DBI-0546594 and an Alfred P. Sloan Research Fellowship.

Appendix A.1: Proof of Lemma 6

Proof By definition, $g_0^*(\mu) = \sup_{x \in \mathbb{R}} (x\mu - C \max(0, x))$. We consider two cases. First, if $\mu < 0$, we have

$$g_0^*(\mu) \geq \sup_{x < 0} (x\mu - C \max(0, x)) = \sup_{x < 0} x\mu = \infty.$$

Therefore, we have $g_0^*(\mu) = \infty$ if $\mu < 0$. Second, if $\mu \geq 0$, we have

$$g_0^*(\mu) = \sup_{x \geq 0} (x\mu - Cx) = \mathbb{I}(\mu \leq C).$$

Putting the above results together, we prove the claim. ■

Appendix A.2: Proof of Lemma 7

Proof The proof has a similar structure as the proof of Lemma 6. Specifically, by definition, the conjugate is

$$g_1^*(\boldsymbol{\mu}) = \sup_{\mathbf{x} \in \mathcal{G}} \left\{ \boldsymbol{\mu}^\top \mathbf{x} - g_1(\mathbf{x}) \right\} = \sup_{\mathbf{x} \in \mathcal{G}} \left\{ \sum_j \mu_j x_j - \max(x_1, \dots, x_L) \right\}.$$

We first show that $\forall i, \mu_i \geq 0$ in order to have finite g_1^* values. Suppose that $\exists j, \mu_j < 0$. Then, we define

$$\mathcal{G}_j = \{\mathbf{x} \in \mathcal{G} : x_j < 0\}, \text{ and } \mathcal{G}_j^o = \{\mathbf{x} \in \mathcal{G} : x_i = 0, \text{ if } i \neq j\}. \quad (47)$$

Since $\mathcal{G}_j^o \subset \mathcal{G}_j \subset \mathcal{G}$, we have

$$g_1^*(\boldsymbol{\mu}) \geq \sup_{\mathbf{x} \in \mathcal{G}_j} \{\boldsymbol{\mu}^\top \mathbf{x} - g_1(\mathbf{x})\} \geq \sup_{\mathbf{x} \in \mathcal{G}_j^o} \{\boldsymbol{\mu}^\top \mathbf{x} - g_1(\mathbf{x})\} = \sup_{x_j \in \mathbb{R}_-} \{x_j \mu_j - 0\} = \infty.$$

Therefore, $g_1^*(\boldsymbol{\mu}) = \infty$ if $\exists j, \mu_j < 0$.

Now, we consider the second case, where $\forall i, \mu_i \geq 0$. We can easily show that

$$\forall \mathbf{x} \in \mathcal{G}, \boldsymbol{\mu}^\top \mathbf{x} - g_1(\mathbf{x}) \leq \sum_i \mu_i g_1(\mathbf{x}) - g_1(\mathbf{x}).$$

Therefore

$$g_1^*(\boldsymbol{\mu}) \leq \sup_{\mathbf{x} \in \mathcal{G}} \left\{ \left(\sum_i \mu_i - C \right) \max(\mathbf{x}) \right\}.$$

Moreover, for any \mathbf{x}_0 that makes $\psi(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \sup_{\mathbf{x} \in \mathcal{G}} \left\{ \left(\sum_i \mu_i - C \right) \max(\mathbf{x}) \right\}$ achieve its supremum, there exists \mathbf{x}' (e.g., $\mathbf{x}'_1 = \max(\mathbf{x}_0)$ and $\mathbf{x}'_j = 0, \forall j \neq 1$), which gives

$$\boldsymbol{\mu}^\top \mathbf{x}' - g_1(\mathbf{x}') = \psi(\boldsymbol{\mu}).$$

Therefore, we have $g_1^*(\boldsymbol{\mu}) \geq \psi(\boldsymbol{\mu})$ and

$$g_1^*(\boldsymbol{\mu}) = \sup_{\mathbf{x} \in \mathcal{G}} \left\{ \left(\sum_i \mu_i - C \right) \max(\mathbf{x}) \right\} = \mathbb{I} \left(\sum_i \mu_i \leq C \right).$$

Putting the above results together proves the claim. ■

Appendix A.3: Proof of Lemma 8

Proof By definition, the conjugate is

$$\begin{aligned} g_2^*(\boldsymbol{\mu}) &= \sup_{\mathbf{x} \in \mathcal{G}'} \left\{ \mu_1 x_1 + \mu_2 x_2 - C \max(0, x_1 - c_1) - C \max(0, x_2 - c_2) \right\}. \\ &= \sup_{x_1 \in \mathbb{R}} \left\{ (\mu_1 - \mu_2) x_1 - C \max(0, x_1 - c_1) - C \max(0, -x_1 - c_2) \right\}. \\ &= \sup_{x_2 \in \mathbb{R}} \left\{ (\mu_2 - \mu_1) x_2 - C \max(0, -x_2 - c_1) - C \max(0, x_2 - c_2) \right\}. \end{aligned}$$

We can see that only the difference $\mu_1 - \mu_2$ or $\mu_2 - \mu_1$ is directly involved in g_2^* . Without loss of generality, we can fix one at 0 and assume the other one is non-negative. Thus, we have $\mu_1 \mu_2 = 0$. Then, we consider two cases.

If $\mu_1 = 0$ (and $\mu_2 \geq 0$), we have

$$\begin{aligned} g_2^*(\boldsymbol{\mu}) &= \sup_{x_2 \in \mathbb{R}} \left\{ \mu_2 x_2 - C \max(0, -x_2 - c_1) - \max(0, x_2 - c_2) \right\} \\ &= c_2 \mu_2 + \sup_{z \in \mathbb{R}} \left\{ \mu_2 z - C \max(0, -z - c_2 - c_1) - \max(0, z) \right\}. \end{aligned}$$

Using the results in the proof of Lemma 6, we can get

$$g_2^*(\boldsymbol{\mu}) = c_2\mu_2 + \mathbb{I}(0 \leq \mu_2 \leq C).$$

Similarly, by symmetry, if $\mu_2 = 0$ (and $\mu_1 \geq 0$), we have

$$g_2^*(\boldsymbol{\mu}) = c_1\mu_1 + \mathbb{I}(0 \leq \mu_1 \leq C).$$

Putting the above results together, we get the conclusions in the lemma. \blacksquare

Appendix A.4: Proof of Lemma 9

Proof Similar structure as the proof of Lemma 10. In this case, the linear expectation operator is $E : \mathcal{P}_{\text{prob}} \rightarrow \mathbb{R}^{2N}$ and the elements of Ep evaluated at the n th example is a two dimensional vector $\boldsymbol{\mu}_n = (\mu_1, \mu_2)$, where $\mu_1 + \mu_2 = 0$ and

$$\mu_1 = \mathbb{E}_{q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta)}[\boldsymbol{\eta}^\top \bar{z}_n]. \quad (48)$$

Then, by the fact that $\max(0, |x| - c) = \max(0, x - c) + \max(0, -x - c)$ and using the g_2 function defined in Lemma 8, we have

$$g(Ep) \stackrel{\text{def}}{=} \mathcal{R}_\epsilon(q(\{\theta_n, z_{nm}, \boldsymbol{\eta}\}|\Theta)) = \sum_n g_2\left(-\boldsymbol{\mu}_n; -y_n + \epsilon, y_n + \epsilon\right).$$

Therefore

$$g^*(\boldsymbol{\omega}) = \sum_n g_2^*(-\boldsymbol{\omega}_n; -y_n + \epsilon, y_n + \epsilon)$$

and

$$g^*(-\boldsymbol{\omega}) = \sum_n g_2^*(\boldsymbol{\omega}_n; -y_n + \epsilon, y_n + \epsilon).$$

By the results in Lemma 5 and Lemma 6, we can derive the conjugate and the optimum solution of \hat{q} . The optimum solution of Θ is due to Lemma 1. Note that the constraints are not directly dependent on Θ . \blacksquare

Appendix A.5: Proof of Lemma 10

Proof By definition, we have $g(Ep) \stackrel{\text{def}}{=} \mathcal{R}_h^c(p(\mathbf{Z}, \boldsymbol{\eta})) = \sum_n g_1(\ell_n^\Delta - Ep(n))$. Let $\boldsymbol{\mu}_n = Ep(n)$. We have the conjugate

$$\begin{aligned} g^*(\boldsymbol{\omega}) &= \sup_{\boldsymbol{\mu}} \left\{ \boldsymbol{\omega}^\top \boldsymbol{\mu} - \sum_n g_1(\ell_n^\Delta - \boldsymbol{\mu}_n) \right\} \\ &= \sum_n \sup_{\boldsymbol{\mu}_n} \left\{ \boldsymbol{\omega}_n^\top \boldsymbol{\mu}_n - g_1(\ell_n^\Delta - \boldsymbol{\mu}_n) \right\} \\ &= \sum_n \sup_{\boldsymbol{\nu}_n} \left\{ \boldsymbol{\omega}_n^\top (\ell_n^\Delta - \boldsymbol{\nu}_n) - g_1(\boldsymbol{\nu}_n) \right\} \\ &= \sum_n \left(\boldsymbol{\omega}_n^\top \ell_n^\Delta + g_1^*(-\boldsymbol{\omega}_n) \right). \end{aligned}$$

Thus,

$$g^*(-\omega) = \sum_n (-\omega_n^\top \ell_n^\Delta + g_1^*(\omega_n)).$$

Using the results of Lemma 5 proves the claim. \blacksquare

Appendix A.6: Proof of Lemma 11

Proof Similar structure as the proof of Lemma 10. In this case, the linear expectation operator is $E : \mathcal{P}_{\text{prob}} \rightarrow \mathbb{R}^{\sum_m |\mathcal{I}_{\text{tr}}^m|}$ and the element of Ep evaluated at the n th example for task m is

$$Ep(n, m) \stackrel{\text{def}}{=} y_{mn} \mathbb{E}_{p(\mathbf{Z}, \boldsymbol{\eta})} [\mathbf{Z} \boldsymbol{\eta}_m]^\top \mathbf{x}_{mn} = \mathbb{E}_{p(\mathbf{Z}, \boldsymbol{\eta})} [y_{mn} (\mathbf{Z} \boldsymbol{\eta}_m)^\top \mathbf{x}_{mn}]. \quad (49)$$

Then, let $g_0 : \mathbb{R} \rightarrow \mathbb{R}$ be a function defined in Lemma 6. We have

$$g(Ep) \stackrel{\text{def}}{=} \mathcal{R}_h^{MT} \left(p(\mathbf{Z}, \boldsymbol{\eta}) \right) = \sum_{m, n \in \mathcal{I}_{\text{tr}}^m} g_0 \left(1 - Ep(n, m) \right).$$

Let $\boldsymbol{\mu} = Ep$. By definition, the conjugate is

$$\begin{aligned} g^*(\omega) &= \sup_{\boldsymbol{\mu}} \left\{ \omega^\top \boldsymbol{\mu} - \sum_{m, n \in \mathcal{I}_{\text{tr}}^m} g_0(1 - \mu_{mn}) \right\} \\ &= \sum_{m, n \in \mathcal{I}_{\text{tr}}^m} \sup_{\mu_{mn}} \left\{ \omega_{mn} \mu_{mn} - g_0(1 - \mu_{mn}) \right\} \\ &= \sum_{m, n \in \mathcal{I}_{\text{tr}}^m} \sup_{\nu_{mn}} \left\{ \omega_{mn} (1 - \nu_{mn}) - g_0(\nu_{mn}) \right\} \\ &= \sum_{m, n \in \mathcal{I}_{\text{tr}}^m} \left(\omega_{mn} + g_0^*(-\omega_{mn}) \right). \end{aligned}$$

Thus,

$$g^*(-\omega) = \sum_{m, n \in \mathcal{I}_{\text{tr}}^m} \left(-\omega_{mn} + g_0^*(\omega_{mn}) \right).$$

By the results in Lemma 5 and Lemma 6, we can derive the conjugate of the problem (41). \blacksquare

Appendix A.7: Inference for MT-iLSVM

In this section, we provide the deviation of the inference algorithm for MT-iLSVM, which is outlined in Algorithm 2 and detailed below.

For MT-iLSVM, the model \mathcal{M} consists of all the latent variables $(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta})$. Let $L_{mn}(p) \stackrel{\text{def}}{=} \mathbb{E}_p[\log p(\mathbf{x}_{mn} | \mathbf{Z}, \mathbf{w}_{mn}, \lambda_{mn}^2)]$ be the expected data likelihood. Then, under the truncated mean-field assumption (44), we have

$$L_{mn}(p) = -\frac{\mathbf{x}_{mn}^\top \mathbf{x}_{mn} - 2\mathbf{x}_{mn}^\top \mathbb{E}_p[\mathbf{Z} \mathbf{w}_{mn}] + \mathbb{E}_p[\mathbf{w}_{mn}^\top \mathbf{U} \mathbf{w}_{mn}]}{2\lambda_{mn}^2} - \frac{D \log(2\pi\lambda_{mn}^2)}{2},$$

where $\mathbf{x}_{mn}^\top \mathbb{E}_p[\mathbf{Z}\mathbf{w}_{mn}] = \sum_k \mathbf{x}_{mn}^\top \boldsymbol{\psi}_{.k}$; $\boldsymbol{\psi}_{.k} \stackrel{\text{def}}{=} (\psi_{1k} \cdots \psi_{Dk})^\top$ is the k th column of $\boldsymbol{\psi} = \mathbb{E}[\mathbf{Z}]$;

$$\mathbb{E}_p[\mathbf{w}_{mn}^\top \mathbf{U} \mathbf{w}_{mn}] = 2 \sum_{j < k} \phi_{mn}^j \phi_{mn}^k \mathbf{U}_{jk} + \sum_k \mathbf{U}_{kk} (K \sigma_{mn}^2 + \Phi_{mn}^\top \Phi_{mn});$$

and $\mathbf{U} \stackrel{\text{def}}{=} \mathbb{E}[\mathbf{Z}^\top \mathbf{Z}]$ is a $K \times K$ matrix, whose element is

$$\mathbf{U}_{ij} = \begin{cases} \sum_d \psi_{di}, & \text{if } i = j \\ \sum_d \psi_{di} \psi_{dj}, & \text{otherwise.} \end{cases}$$

For the KL-divergence term, we have $\text{KL}(p(\mathcal{M}) \parallel \pi(\mathcal{M})) = \text{KL}(p(\boldsymbol{\nu}) \parallel \pi(\boldsymbol{\nu})) + \text{KL}(p(\mathbf{W}) \parallel \pi(\mathbf{W})) + \text{KL}(p(\mathbf{Z}) \parallel \pi(\mathbf{Z})) + \text{KL}(p(\boldsymbol{\eta}) \parallel \pi(\boldsymbol{\eta}))$, where the individual terms are

$$\begin{aligned} \text{KL}(p(\boldsymbol{\nu}) \parallel \pi(\boldsymbol{\nu})) &= \sum_{k=1}^K \left((\gamma_{k1} - \alpha)(\psi(\gamma_{k1}) - \psi(\gamma_{k1} + \gamma_{k2})) + (\gamma_{k2} - 1)(\psi(\gamma_{k2}) - \psi(\gamma_{k1} + \gamma_{k2})) \right. \\ &\quad \left. - \log \frac{\Gamma(\gamma_{k1})\Gamma(\gamma_{k2})}{\Gamma(\gamma_{k1} + \gamma_{k2})} \right) - K \log \alpha, \\ \text{KL}(p(\mathbf{Z}) \parallel \pi(\mathbf{Z})) &= \sum_{dk} \left(-\psi_{dk} \sum_{j=1}^k \mathbb{E}_p[\log \nu_j] - (1 - \psi_{dk}) \mathbb{E}_p[\log(1 - \prod_{j=1}^k \nu_j)] \right. \\ &\quad \left. + \psi_{dk} \log \psi_{dk} + (1 - \psi_{dk}) \log(1 - \psi_{dk}) \right) \\ \text{KL}(p(\mathbf{W}) \parallel \pi(\mathbf{W})) &= \sum_{mn} \left(\frac{K \sigma_{mn}^2 + \Phi_{mn}^\top \Phi_{mn}}{2 \sigma_{m0}^2} - \frac{K(1 + \log \frac{\sigma_{mn}^2}{\sigma_{m0}^2})}{2} \right). \end{aligned}$$

where $\psi(\cdot)$ is the digamma function and $\mathbb{E}_p[\log \nu_j] = \psi(\gamma_{j1}) - \psi(\gamma_{j1} + \gamma_{j2})$. For $\text{KL}(p(\boldsymbol{\eta}) \parallel \pi(\boldsymbol{\eta}))$, we do not need to write it explicitly, as we shall see. Finally, the effective discriminant function is

$$f_m(\mathbf{x}_{mn}; p(\mathbf{Z}, \boldsymbol{\eta})) = \boldsymbol{\eta}_m^\top \boldsymbol{\psi}^\top \mathbf{x}_{mn} = \sum_{k=1}^K \mathbb{E}_p[\eta_{mk}] \boldsymbol{\psi}_{.k}^\top \mathbf{x}_{mn}.$$

All the above terms can be easily computed, except the term $\mathbb{E}_p[\log(1 - \prod_{j=1}^k \nu_j)]$. Here, we adopt the multivariate lower bound (Doshi-Velez et al., 2009)

$$\begin{aligned} \mathbb{E}_p[\log(1 - \prod_{j=1}^k \nu_j)] &\geq \sum_{m=1}^k q_{km} \psi(\gamma_{m2}) + \sum_{m=1}^{k-1} \left(\sum_{n=m+1}^k q_{kn} \right) \psi(\gamma_{m1}) \\ &\quad - \sum_{m=1}^k \left(\sum_{n=m}^k q_{kn} \right) \psi(\gamma_{m1} + \gamma_{m2}) + \mathcal{H}(q_{k.}), \end{aligned}$$

where the variational parameters $q_{k.} = (q_{k1} \cdots q_{kk})^\top$ belong to the k -simplex, and $\mathcal{H}(q_{k.})$ is the entropy of $q_{k.}$. The tightest lower bound is achieved by setting $q_{k.}$ to be the optimum value

$$q_{km} = \frac{1}{Z_k} \exp \left(\psi(\gamma_{m2}) + \sum_{n=1}^{m-1} \psi(\gamma_{n1}) - \sum_{n=1}^m \psi(\gamma_{n1} + \gamma_{n2}) \right), \quad (50)$$

Algorithm 2 Inference Algorithm of MT-iLSVM

- 1: **Input:** data $\mathcal{D} = \{(\mathbf{x}_{mn}, y_{mn})\}_{m,n \in \mathcal{I}_{\text{tr}}} \cup \{\mathbf{x}_{mn}\}_{m,n \in \mathcal{I}_{\text{tst}}}$, constants α and C
 - 2: **Output:** distributions $p(\boldsymbol{\nu})$, $p(\mathbf{Z})$, $p(\mathbf{W})$, $p(\boldsymbol{\eta})$ and hyper-parameters σ_{m0}^2 and λ_{mn}^2
 - 3: Initialize $\gamma_{k1} = \alpha$, $\gamma_{k2} = 1$, $\psi_{dk} = 0.5 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.001)$, $\Phi_{mn} = 0$, $\sigma_{mn}^2 = \sigma_{m0}^2 = 1$, $\boldsymbol{\mu}_m = 0$, λ_{mn}^2 is computed from \mathcal{D} .
 - 4: **repeat**
 - 5: **repeat**
 - 6: update $(\gamma_{k1}, \gamma_{k2})$ using Eq. (52), $\forall 1 \leq k \leq K$;
 - 7: update ϕ_{mn}^k and σ_{mn}^2 using Eq. (51), $\forall m, \forall n, \forall 1 \leq k \leq K$;
 - 8: update ψ_{dk} using Eq. (53), $\forall 1 \leq d \leq D, \forall 1 \leq k \leq K$;
 - 9: **until** relative change of L is less than τ (e.g., $1e^{-3}$) or iteration number is T (e.g., 10)
 - 10: **for** $m = 1$ **to** M **do**
 - 11: solve the dual problem (54) using a binary SVM learner.
 - 12: **end for**
 - 13: update the hyper-parameters σ_{m0}^2 using Eq. (55) and λ_{mn}^2 using Eq. (56). (*Optional*)
 - 14: **until** relative change of L is less than τ' (e.g., $1e^{-4}$) or iteration number is T' (e.g., 20)
-

where Z_k is a normalization factor to make q_k be a distribution. We denote the tightest lower bound by \mathcal{L}_k^ν . Replacing the term $\mathbb{E}_p[\log(1 - \prod_{j=1}^k \nu_j)]$ with its lower bound \mathcal{L}_k^ν , we can have an upper bound of $\text{KL}(p(\mathcal{M}) \parallel \pi(\mathcal{M}))$ and we denote this upper bound by $\mathcal{L}(p)$.

With the above terms and the upper bound $\mathcal{L}(p)$, we can implement the general procedure outlined in Algorithm 1 to solve the MT-iLSVM problem. Specifically, the inference procedure iteratively solves the following steps, as summarized in Algorithm 2:

Infer $p(\boldsymbol{\nu})$, $p(\mathbf{Z})$ and $p(\mathbf{W})$: For $p(\mathbf{W})$, since both the prior $\pi(\mathbf{W})$ and $p(\mathbf{W})$ are Gaussian, we can easily derive the update rules, similar as in Gaussian mixture models

$$\begin{aligned} \phi_{mn}^k &= \frac{\sum_d x_{mn}^d \psi_{dk} - \sum_{j \neq k} \phi_{mn}^j \mathbf{U}_{kj}}{\lambda_{mn}^2} \left(\frac{1}{\sigma_{m0}^2} + \frac{\sum_d \psi_{dk}}{\lambda_{mn}^2} \right)^{-1} \\ \sigma_{mn}^2 &= \left(\frac{1}{\sigma_{m0}^2} + \frac{1}{K} \sum_k \frac{\mathbf{U}_{kk}}{\lambda_{mn}^2} \right)^{-1} \end{aligned} \quad (51)$$

For $p(\boldsymbol{\nu})$, we have the update rules similar as in (Doshi-Velez et al., 2009), that is,

$$\begin{aligned} \gamma_{k1} &= \alpha + \sum_{m=k}^K \sum_{d=1}^D \psi_{dm} + \sum_{m=k+1}^K (D - \sum_{d=1}^D \psi_{dm}) \left(\sum_{i=k+1}^m q_{mi} \right) \\ \gamma_{k2} &= 1 + \sum_{m=k}^K (D - \sum_{d=1}^D \psi_{dm}) q_{mk}. \end{aligned} \quad (52)$$

For $p(\mathbf{Z})$, we have the mean-field update equation as

$$\psi_{dk} = \frac{1}{1 + e^{-\vartheta_{dk}}}, \quad (53)$$

where

$$\begin{aligned} \vartheta_{dk} = & \sum_{j=1}^k \mathbb{E}_p[\log v_j] - \mathcal{L}_k^\nu - \sum_{mn} \frac{1}{2\lambda_{mn}^2} \left((K\sigma_{mn}^2 + (\phi_{mn}^k)^2) \right. \\ & \left. - 2x_{mn}^d \phi_{mn}^k + 2 \sum_{j \neq k} \phi_{mn}^j \phi_{mn}^k \psi_{dj} \right) + \sum_{m,n \in \mathcal{I}_{tr}^m} y_{mn} \mathbb{E}_p[\eta_{mk}] x_{mn}^d. \end{aligned}$$

Infer $p(\boldsymbol{\eta})$ and solve for $\boldsymbol{\omega}$: By the convex duality theory, we have the solution

$$\begin{aligned} p(\boldsymbol{\eta}) & \propto \pi(\boldsymbol{\eta}) \exp \left\{ \sum_{m,n \in \mathcal{I}_{tr}^m} y_{mn} \omega_{mn} \boldsymbol{\eta}_m^\top \boldsymbol{\psi}^\top \mathbf{x}_{mn} \right\} \\ & = \prod_{m=1}^M \pi(\boldsymbol{\eta}_m) \exp \left\{ \boldsymbol{\eta}_m^\top \left(\sum_{n \in \mathcal{I}_{tr}^m} y_{mn} \omega_{mn} \boldsymbol{\psi}^\top \mathbf{x}_{mn} \right) \right\}. \end{aligned}$$

Therefore, we can see that although we did not assume $p(\boldsymbol{\eta})$ is factorized, we can get the induced factorization form $p(\boldsymbol{\eta}) = \prod_m p(\boldsymbol{\eta}_m)$, where

$$p(\boldsymbol{\eta}_m) \propto \pi(\boldsymbol{\eta}_m) \exp \left\{ \boldsymbol{\eta}_m^\top \left(\sum_{n \in \mathcal{I}_{tr}^m} y_{mn} \omega_{mn} \boldsymbol{\psi}^\top \mathbf{x}_{mn} \right) \right\}.$$

Here, we assume $\pi(\boldsymbol{\eta}_m)$ is standard normal. Then, we have $p(\boldsymbol{\eta}_m) = \mathcal{N}(\boldsymbol{\eta}_m | \boldsymbol{\mu}_m, I)$, where

$$\boldsymbol{\mu}_m = \sum_{n \in \mathcal{I}_{tr}^m} y_{mn} \omega_{mn} \boldsymbol{\psi}^\top \mathbf{x}_{mn}.$$

The optimum dual parameters can be obtained by solving the following M independent dual problems

$$\max_{\boldsymbol{\omega}_m} -\frac{1}{2} \boldsymbol{\mu}_m^\top \boldsymbol{\mu}_m + \sum_{n \in \mathcal{I}_{tr}^m} \omega_{mn} \quad \text{s.t.} : 0 \leq \omega_{mn} \leq 1, \forall n \in \mathcal{I}_{tr}^m, \quad (54)$$

which (and its primal form) can be efficiently solved with a binary SVM solver, such as SVM-light.

As we have stated, the hyperparameters σ_0^2 and λ_{mn}^2 can be set a priori or estimated from the data. The empirical estimation can be easily done with closed form solutions. For MT-iLSVM, we have

$$\sigma_{m0}^2 = \frac{\sum_{n=1}^{N_m} (K\sigma_{mn}^2 + \Phi_{mn}^\top \Phi_{mn})}{KN_m} \quad (55)$$

$$\lambda_{mn}^2 = \frac{\mathbf{x}_{mn}^\top \mathbf{x}_{mn} - 2\mathbf{x}_{mn}^\top \mathbb{E}_p[\mathbf{Z}\mathbf{w}_{mn}] + \mathbb{E}_p[\mathbf{w}_{mn}^\top \mathbf{U}\mathbf{w}_{mn}]}{D}. \quad (56)$$

Appendix A.8: Inference for Infinite Latent SVM

In this section, we develop the inference algorithm for iLSVM based on the stick-breaking construction of the IBP prior. The algorithm is outlined in Algorithm 3.

Similar as in the inference for MT-iLSVM, we make the additional constraint about the feasible distribution

$$p(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta}) = p(\boldsymbol{\eta})p(\mathbf{W}|\Phi, \Sigma) \prod_n \left(\prod_{k=1}^K p(z_{nk}|\psi_{nk}) \right) \prod_{k=1}^K p(\nu_k|\gamma_k),$$

where K is the truncation level; $p(\mathbf{W}|\Phi, \Sigma) = \prod_k \mathcal{N}(\mathbf{W}_{.k}|\Phi_{.k}, \sigma_k^2 I)$; $p(z_{nk}|\phi_{nk}) = \text{Bernoulli}(\phi_{nk})$; and $p(\nu_k|\gamma_k) = \text{Beta}(\gamma_{k1}, \gamma_{k2})$. Then, we solve the unconstrained problem using convex duality with dual parameters being $\boldsymbol{\omega}$. Let $L_n(p) \stackrel{\text{def}}{=} \mathbb{E}_p[\log p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{W})]$. We have

$$L_n(p) = -\frac{\mathbf{x}_n^\top \mathbf{x}_n - 2\mathbf{x}_n^\top \Phi \mathbb{E}_p[\mathbf{z}_n]^\top + \mathbb{E}_p[\mathbf{z}_n \mathbf{A} \mathbf{z}_n^\top]}{2\sigma_{n0}^2} - \frac{D \log(2\pi\sigma_{n0}^2)}{2}, \quad (57)$$

where $\mathbf{A} \stackrel{\text{def}}{=} \mathbb{E}_p[\mathbf{W}^\top \mathbf{W}]$ is a $K \times K$ matrix; $\mathbf{x}_n^\top \Phi \mathbb{E}_p[\mathbf{z}_n]^\top = 2 \sum_k \psi_{nk}(\mathbf{x}_n^\top \Phi_{.k})$; and

$$\mathbb{E}_p[\mathbf{z}_n \mathbf{A} \mathbf{z}_n^\top] = 2 \sum_{j < k} \psi_{nj} \psi_{nk} \mathbf{A}_{jk} + \sum_k \psi_{nk} (D\sigma_k^2 + \mathbf{A}_{kk}).$$

The effective discriminant function is $f(y, \mathbf{x}_n) = \sum_k \mathbb{E}_p[\eta_y^k] \psi_{nk}$. Again, for computational tractability, we need the lower bound \mathcal{L}_k^ν of the term $\mathbb{E}_p[\log(1 - \prod_{j=1}^k v_j)]$. Using this lower bound, we can get an upper bound of the KL-divergence term. Then, the inference procedure iteratively solves the following steps:

Infer $p(\boldsymbol{\nu})$, $p(\mathbf{Z})$ and $p(\mathbf{W})$: For $p(\mathbf{W})$, we have the update rules

$$\begin{aligned} \Phi_{.k} &= \sum_n \frac{\psi_{nk}}{\sigma_{n0}^2} \left(\mathbf{x}_n - \sum_{j \neq k} \psi_{nj} \Phi_{.j} \right) \left(1 + \sum_n \frac{\psi_{nk}}{\sigma_{n0}^2} \right)^{-1} \\ \sigma_k^2 &= \left(1 + \sum_n \frac{\psi_{nk}}{\sigma_{n0}^2} \right)^{-1}. \end{aligned} \quad (58)$$

For $p(\boldsymbol{\nu})$, we have the update rules similar as in (Doshi-Velez et al., 2009), that is,

$$\begin{aligned} \gamma_{k1} &= \alpha + \sum_{m=k}^K \sum_{n=1}^N \psi_{nm} + \sum_{m=k+1}^K (N - \sum_{n=1}^N \psi_{nm}) \left(\sum_{i=k+1}^m q_{mi} \right) \\ \gamma_{k2} &= 1 + \sum_{m=k}^K (N - \sum_{n=1}^N \psi_{nm}) q_{mk}, \end{aligned} \quad (59)$$

where $q_{.k}$ is computed in the same way as in Eq. (50). For $p(\mathbf{Z})$, the mean-field update equation for ψ is

$$\psi_{nk} = \frac{1}{1 + e^{-\vartheta_{nk}}}, \quad (60)$$

where

$$\begin{aligned} \vartheta_{nk} &= \sum_{j=1}^k \mathbb{E}_p[\log v_j] - \mathcal{L}_k^\nu(p) - \frac{1}{2\sigma_{n0}^2} (D\sigma_k^2 + \Phi_{.k}^\top \Phi_{.k}) \\ &\quad + \frac{1}{\sigma_{n0}^2} \Phi_{.k}^\top \left(\mathbf{x}_n - \sum_{j \neq k} \psi_{nj} \Phi_{.j} \right) + \sum_y \omega_n^y \mathbb{E}_p[\eta_{y_n}^k - \eta_y^k]. \end{aligned}$$

Algorithm 3 Inference Algorithm of iLSVM

- 1: **Input:** data $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n \in \mathcal{I}_{\text{tr}}} \cup \{\mathbf{x}_n\}_{n \in \mathcal{I}_{\text{tst}}}$, constants α and C
 - 2: **Output:** distributions $p(\boldsymbol{\nu})$, $p(\mathbf{Z})$, $p(\mathbf{W})$, $p(\boldsymbol{\eta})$ and hyper-parameters σ_0^2 and σ_{n0}^2
 - 3: Initialize $\gamma_{k1} = \alpha$, $\gamma_{k2} = 1$, $\psi_{nk} = 0.5 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.001)$, $\Phi_{\cdot k} = 0$, $\sigma_k^2 = \sigma_0^2 = 1$, $\boldsymbol{\mu} = 0$, σ_{n0}^2 is computed from \mathcal{D} .
 - 4: **repeat**
 - 5: **repeat**
 - 6: update $(\gamma_{k1}, \gamma_{k2})$ using Eq. (59), $\forall 1 \leq k \leq K$;
 - 7: update $\Phi_{\cdot k}$ and σ_k^2 using Eq. (58), $\forall 1 \leq k \leq K$;
 - 8: update ψ_{nk} using Eq. (60), $\forall n \in \mathcal{I}_{\text{tr}}, \forall 1 \leq k \leq K$;
 - 9: update ψ_{nk} using Eq. (60), but ϑ_{nk} doesn't have the last term, $\forall n \in \mathcal{I}_{\text{tst}}, \forall 1 \leq k \leq K$;
 - 10: **until** relative change of L is less than τ (e.g., $1e^{-3}$) or iteration number is T (e.g., 10)
 - 11: solve the dual problem (61) (or its primal form) using a multi-class SVM learner.
 - 12: update the hyper-parameters σ_0^2 using Eq. (62) and σ_{n0}^2 using Eq. (63). (*Optional*)
 - 13: **until** relative change of L is less than τ' (e.g., $1e^{-4}$) or iteration number is T' (e.g., 20)
-

For testing data, ϑ_{nk} does not have the last term because of the absence of large-margin constraints.

Infer $p(\boldsymbol{\eta})$ and solve for $\boldsymbol{\omega}$: By the convex duality theory, we have

$$p(\boldsymbol{\eta}) \propto \pi(\boldsymbol{\eta}) \exp \left\{ \boldsymbol{\eta}^\top \left(\sum_{n \in \mathcal{I}_{\text{tr}}} \sum_y \omega_n^y \mathbb{E}_p[\mathbf{g}(y_n, \mathbf{x}_n, \mathbf{z}_n) - \mathbf{g}(y, \mathbf{x}_n, \mathbf{z}_n)] \right) \right\}.$$

For the standard normal prior $\pi(\boldsymbol{\eta})$, we have that $q(\boldsymbol{\eta})$ is also normal, with mean

$$\boldsymbol{\mu} = \sum_{n \in \mathcal{I}_{\text{tr}}} \sum_y \omega_n^y \mathbb{E}_p[\mathbf{g}(y_n, \mathbf{x}_n, \mathbf{z}_n) - \mathbf{g}(y, \mathbf{x}_n, \mathbf{z}_n)]$$

and identity covariance matrix. The dual problem is

$$\max_{\boldsymbol{\omega}} \quad -\frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\mu} + \sum_{n \in \mathcal{I}_{\text{tr}}} \sum_y \omega_n^y \quad \text{s.t.} : \quad 0 \leq \sum_y \omega_n^y \leq C, \forall n \in \mathcal{I}_{\text{tr}}, \quad (61)$$

which (and its primal form) can be efficiently solved with a multi-class SVM solver.

Similar as in MT-iLSVM, the hyperparameters σ_0^2 and σ_{n0}^2 can be set a priori or estimated from the data. The empirical estimation can be easily done with closed form solutions. For iLSVM, we have

$$\sigma_0^2 = \frac{\sum_{k=1}^K (D \sigma_k^2 + \Phi_{\cdot k}^\top \Phi_{\cdot k})}{KD} \quad (62)$$

$$\sigma_{n0}^2 = \frac{\mathbf{x}_n^\top \mathbf{x}_n - 2 \mathbf{x}_n^\top \Phi \mathbb{E}_p[\mathbf{z}_n]^\top + \mathbb{E}_p[\mathbf{z}_n \mathbf{A} \mathbf{z}_n^\top]}{D}. \quad (63)$$

References

- Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *International Conference on Learning Theory*, 2006.
- Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, (6):1817–1853, 2005.
- Charles E. Antoniak. Mixture of Dirichlet process with applications to Bayesian nonparametric problems. *Annals of Statistics*, (273):1152–1174, 1974.
- Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. In *Advances in Neural Information Processing Systems*, 2007.
- Bart Bakker and Tom Heskes. Task clustering and gating for Bayesian multitask learning. *Journal of Machine Learning Research*, (4):83–99, 2003.
- Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, 2002.
- Kedar Bellare, Gregory Druck, and Andrew McCallum. Alternating projections for learning with expectation constraints. In *Uncertainty in Artificial Intelligence*, 2009.
- Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, 2006.
- David Blei and Peter Frazier. Distance dependent Chinese restaurant process. In *International Conference on Machine Learning*, 2010.
- David Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, (3):993–1022, 2003.
- Edwin Bonilla, Kian Ming Chai, and Christopher Williams. Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems*, 2008.
- Jonathan Borwein and Qiji Zhu. *Techniques of Variational Analysis: An Introduction*. Springer, New York, NY, 2005.
- Stephen Boyd and Lieven Vandenbergh. *Convex Optimization*. Cambridge University Press, 2004.
- Ning Chen, Jun Zhu, and Fuchun Sun. Infinite exponentail family harmoniums. In *NIPS Workshop on Bayesian Nonparametric Methods: Hope or Hype?*, 2011.
- Ning Chen, Jun Zhu, and Eric P. Xing. Predictive subspace learning for multiview data: a large margin approach. In *Advances in Neural Information Processing Systems*, 2010.
- Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

- Finale Doshi-Velez, Kurt Miller, Jurgen Van Gael, and Yee Whye Teh. Variational inference for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, 2009.
- Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, (8):1217–1260, 2007.
- David Dunson and Shyamal Peddada. Bayesian nonparametric inferences on stochastic ordering. *ISDS Discussion Paper*, 2, 2007.
- Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(24):429–433, 2002.
- Thomas Ferguson. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 1973.
- Morten Frydenberg. The chain graph markov property. *Scandinavian Journal of Statistics*, 17:333–353, 1990.
- Kuzman Ganchev, João. Graca, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, (11):2001–2094, 2010.
- Samuel Gershman and David Blei. A tutorial on bayesian nonparametric models. *arXiv:1106.2697v2*, 2011.
- João Graca, Kuzman Ganchev, Ben Taskar, and Fernando Pereira. Posterior vs. parameter sparsity in latent variable models. In *Advances in Neural Information Processing Systems*, 2009.
- Thomas Griffiths and Zoubin Ghahramani. Infinite latent feature models and the Indian buffet process. Technical report, University College London, GCNU TR2005-001, 2005.
- Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 2004.
- Peter D. Hoff. Bayesian methods for partial stochastic orderings. *Biometrika*, 90:303–317, 2003.
- Seungil Huh and Stephen Fienberg. Discriminative topic modeling based on manifold learning. In *KDD*, 2010.
- Kazufumi Ito and Karl Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*. Advances in Design and Control, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.
- Tommi Jaakkola, Meila Meila, and Tony Jebara. Maximum entropy discrimination. In *Advances in Neural Information Processing Systems*, 1999.

- Tony Jebara. *Discriminative, Generative and Imitative Learning*. PhD thesis, Media Laboratory, MIT, Dec 2001.
- Tony Jebara. Multitask sparsity via maximum entropy discrimination. *Journal of Machine Learning Research*, (12):75–110, 2011.
- Thorsten Joachims. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning*, 1999.
- Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. *An introduction to variational methods for graphical models*. M. I. Jordan (Ed.), Learning in Graphical Models, Cambridge: MIT Press, Cambridge, MA, 1999.
- Mohammad E. Khan, Guillaume Bouchard, Benjamin Marlin, and Kevin Murphy. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Processing Systems*, 2010.
- David C. Knill and Alexandre Pouget. The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neuroscience*, 27(12):712–719, 2004.
- John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, 2001.
- Percy Liang, Michael Jordan, and Dan Klein. Learning from measurements in exponential families. In *International Conference on Machine Learning*, 2009.
- Steven N. MacEachern. Dependent nonparametric process. In *the Section on Bayesian Statistical Science of ASA*, 1999.
- Thomas L. Magnanti. Fenchel and lagrange duality are equivalent. *Mathematical Programming*, (7):253–258, 1974.
- Gideon Mann and Andrew McCallum. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, (11):955–984, 2010.
- Kurt Miller, Thomas Griffiths, and Michael Jordan. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*, 2009.
- Iain Murray and Zoubin Ghahramani. Bayesian learning in undirected graphical models: Approximate MCMC algorithms. In *Uncertainty in Artificial Intelligence*, 2004.
- Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1988.
- Yuan (Alan) Qi, Martin Szummer, and Thomas P. Minka. Bayesian conditional random fields. In *International Conference on Artificial Intelligence and Statistics*, 2005.
- Piyush Rai and Hal Daume III. Infinite predictor subspace models for multitask learning. In *International Conference on Artificial Intelligence and Statistics*, 2010.

- Charles E. Rasmussen and Zoubin Ghahramani. Infinite mixtures of Gaussian process experts. In *Advances in Neural Information Processing Systems*, 2002.
- Edward Schofield. *Fitting maximum-entropy models on large sample spaces*. PhD thesis, PhD thesis, Department of Computing, Imperial College London, 2006.
- Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin Markov networks. In *Advances in Neural Information Processing Systems*, 2003.
- Yee Whye Teh, Dilan Görür, and Zoubin Ghahramani. Stick-breaking construction of the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, 2007.
- Yee Whye Teh, Michael Jordan, Matthew Beal, and David Blei. Hierarchical Dirichlet process. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
- Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.
- V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.
- Max Welling and Sridevi Parise. Bayesian random fields: The bethe-laplace approximation. In *Uncertainty in Artificial Intelligence*, 2006.
- Max Welling, Ian Porteous, and Kenichi Kurihara. Exchangeable inconsistent priors for bayesian posterior inference. In *Workshop on Information Theory and Applications*, 2012.
- Max Welling, M. Rosen-Zvi, and G. Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems*, 2004.
- Sinead Williamson, Peter Orbanz, and Zoubin Ghahramani. Dependent indian buffet processes. In *International Conference on Artificial Intelligence and Statistics*, 2010.
- Ya Xue, David Dunson, and Lawrence Carin. The matrix stick-breaking process for flexible multi-task learning. In *International Conference on Machine Learning*, 2007.
- Arnold Zellner. Optimal information processing and Bayes’ theorem. *The American Statistician*, 42:278–280, 1988.
- Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. In *Uncertainty in Artificial Intelligence*, 2010.
- Jun Zhu, Amir Ahmed, and Eric P. Xing. MedLDA: Maximum margin supervised topic models for regression and classification. In *International Conference on Machine Learning*, 2009.
- Jun Zhu, Ning Chen, and Eric P. Xing. Infinite latent SVM for classification and multi-task learning. In *Advances in Neural Information Processing Systems*, 2011a.

Jun Zhu, Ning Chen, and Eric P. Xing. Infinite SVM: a Dirichlet process mixture of large-margin kernel machines. In *International Conference on Machine Learning*, 2011b.

Jun Zhu and Eric P. Xing. Maximum entropy discrimination Markov networks. *Journal of Machine Learning Research*, (10):2531–2569, 2009.